Etičko razmatranje moralnog statusa umjetno inteligentnih sustava

Kljajić, Filipa

Sveučilište u Zadru

Odjel za filozofiju

Diplomski sveučilišni studij filozofije; smjer: nastavnički (dvopredmetni)

Filipa Kljajić

Etičko razmatranje moralnog statusa umjetno inteligentnih sustava

Diplomski rad

Zadar, 2019.

Sveučilište u Zadru

Odjel za filozofiju
Diplomski sveučilišni studij filozofije; smjer: nastavnički (dvopredmetni)

Etičko razmatranje moralnog statusa umjetno inteligentnih sustava

Diplomski rad

Studentica:
Filipa Kljajić

Mentorica:
prof. dr. sc. Iris Tićac

Zadar, 2019.

Izjava o akademskoj čestitosti

Ja, Filipa Kljajić**,** ovime izjavljujem da je moj diplomski rad pod naslovom *Etičko razmatranje moralnog statusa umjetno inteligentnih sustava* rezultat mojega vlastitog rada, da se temelji na mojim istraživanjima te da se oslanja na izvore i radove navedene u bilješkama i popisu literature. Ni jedan dio mojega rada nije napisan na nedopušten način, odnosno nije prepisan iz necitiranih radova i ne krši bilo čija autorska prava.

Izjavljujem da ni jedan dio ovoga rada nije iskorišten u kojem drugom radu pri bilo kojoj drugoj visokoškolskoj, znanstvenoj, obrazovnoj ili inoj ustanovi.

Sadržaj mojega rada u potpunosti odgovara sadržaju obranjenoga i nakon obrane uređenoga rada.

Zadar, 24. travnja 2019.

SADRŽAJ

UVOD

U današnje doba svjedočimo procvatu tehnologije, specifično, nevjerojatno brzom napretku u području umjetno inteligentnih sustava. O odnosu čovjeka i umjetne inteligencije postoje danas snažna divergentna shvaćanja. Čovjek nije sasvim svjestan koliko se oslanja na umjetno inteligentne sustave u svakodnevnom životu. Ovi su sustavi poprilično zaslužni za olakšavanje i poboljšanje mnogih dijelova našeg života. Veoma lagano i brzo možemo obaviti određene jednostavne stvari poput kupovine, pronalaženja relevantnih informacija, traženja poslova i sl. Iako su napredak, takvi sustavi imaju puno veći utjecaj na ljudski život od jednostavne uštede čovjekova vremena. Sustavi umjetne inteligencije prate svaki naš pokret prilagođavajući se našim potrebama i savjetujući nas što kupiti, što jesti, kuda se kretati, što vidjeti. Ovi sustavi također prate medicinske i općeznanstvene trendove pa čak i ekonomska kretanja na globalnom tržištu, stoga nije čudo da stručnjaci u ovim poljima danas redovito koriste umjetne inteligentne sustave u svakodnevnom radu. Iz navedenih razloga neki su mišljenja da nevjerojatna preciznost, brzina i sposobnost prikupljanja informacija i učenja čini umjetne sustave kompetentnijima od čovjeka u mnogim područjima. Prema drugima načelno je nemoguće da umjetni sustavi ikada nadomjeste čovjeka, posebice kada se radi o pravnim, pedagoškim i sličnim zadaćama. U svakom slučaju razvoj umjetnih sustava ima utjecaja na samorazumijevanje čovjeka i njegova svijeta.

Istraživanje umjetne inteligencije mlada je znanost i najčešće se propituje u području informatike i prirodoznanstvenih disciplina. No, pitanje kako se čovjek razlikuje od umjetne inteligencije nije empirijsko, nego filozofijsko pitanje, stoga je potrebna filozofijska refleksija ovog fenomena. Ona je potrebna tim više što su raširena shvaćanja prema kojima bi se u svrhu poboljšanja kvalitete života ili što učinkovitijeg rješavanja problema na koje nailazimo, razvoj ovog tehnološkog fenomena trebao usmjeriti ka stvaranju opće umjetne inteligencije za koju određeni teoretičari smatraju da će funkcionirati poput čovjekove generalne inteligencije. Krajnji je cilj načiniti i super-inteligentni umjetni sustavkoji bi nadmašio čovjeka, i čiji bi izum iz temelja izmijenio našu civilizaciju.

Pitanja i problem vezani uzutjecaj razvoja umjetne inteligencije na čovjeka i društvo predmet je mnogih rasprava, no u ovom se radu problematizira pitanje trebamo li se moralno odnositi prema sustavima umjetne inteligencije, tj. trebamo li ju smatrati moralnim pacijentom – bićem koje je objekt moralnog djelovanjano samo ne može djelovati na način koji bi

okarakterizirali kao (ne)moralan[1] te pod kojim je uvjetima moguće smatrati umjetni sustav moralnim pacijentom.

Navedeno se pitanje može okarakterizirati kao zanimljivo jer se relativno malo stručnjaka bavi etičnošću ovog odnosa. Generalno želimo i trebamo znati što trebamo uključiti u krug etičkog razmatranja jer ne želimo nanositi zlo već biti moralni prema onima (ili onome) koji to zaslužuju. Stoga, ako želimo biti pravedni i ne činiti nepravdu zbog našeg neznanja kao što smo to radili u prošlosti ljudima, životinjama i prirodi, trebamo logički razmisliti ima li smisla i umjetne sustave tretirati moralno. Raditi istu grešku kao i u prošlosti zbog predrasuda i neznanja više nije opravdano.

Kako bismo pokušali odgovoriti na pitanje li ima smisla davati moralni značaj umjetno inteligentnom sustavu, za početak će u radu dati informacije vezane za UI tehnologiju koje su relevantne za ovu raspravu, potom će se etičko-argumentiranom raspravom odgovoriti na tvrdnje kojima se govori da je besmisleno pripisivati sustavima umjetne inteligencije moralni status. U radu će se također tematizirati pitanje svijesti računalnih sustava. Nakon toga će se uz pomoć dviju etičkih teorija – deontologije i utilitarizma, pokušati pokazati da, donekle, možemo dati umjetno inteligentnim sustavima moralni status.

Rad završava problematikom umjetnog inteligentnog sustava koji bi imao iskustvo boli jer dokaz da umjetni sustav osjeti bol, kao i to da posjeduje svijest, se najčešće postavlja kao uvjet davanja statusa moralnog pacijenta umjetno inteligentnom sustavu.

[1] Moralnog djelatelja možemo razumjeti kao biće čije djelovanje možemo opisati kao moralno ili nemoralno, no na koje se također može isto tako i djelovati, znači čovjeka shvaćamo kao moralnog agenta i pacijenta dok druga bića i djecu samo kao moralne pacijente.
Vidjeti šire: Tom Regan, T*he Case for Animal Rights*, University of California Press, 2004., str. 151- 152.

1. UMJETNA INTELIGENCIJA (UI)

Postoje razne kolokvijalne definicije umjetne inteligencije te se prema njima obično misli na mehanizme ili strojeve koji oponašaju ljudske kognitivne funkcije. No u računalnim znanostima se obično koristi definicija Davida Poolea i Alana K. Mackworthova[2] koja opisuje umjetnu inteligenciju ili UI kao područje znanosti koje se bavi proučavanjem računalnog agenta koji se ponaša inteligentno. Cilj istraživanja i gradnjeumjetne inteligencije jest dizajniranje korisne i inteligentne naprave.

Prema njihovoj definiciji agent (djelatelj) može biti sve što djeluje u svijetu, stoga se ova široka definicija može odnositi na sva živa bića, ali i na stvari poput predmeta i robota .Prema njima agent je inteligentan ako:

– čini ono što je prikladno u danoj situaciji kako bi postigao cilj

– može se prilagoditi situaciji i promjeni ciljeva

– uči na osnovi vlastitog iskustva

–opredjeljuje se za one izbore koje može ostvariti s obzirom na vlastita ograničenja.

Računalni agent je agent čije se odluke o vlastitim akcijama mogu objasniti računalnim radnjama. Drugim riječima, odluke takvog agenta mogu se svesti na primitivne operacije koje se mogu implementirati u hardver. Znači, kada se misli na inteligentnog agenta, pojam inteligencije se odnosi na onu vrstu inteligencije koja vodi do boljeg performansa, a ne na inteligentnu misao.

Od američkog filozofa John R. Searla potječe podjela na „jake“ i „slabe“ teorije, tj. shvaćanja umjetne inteligencije. Prema „slaboj ili opreznoj“ teoriji umjetni sustavi, posebice kompjutori samo su pmoćno sredstvo u istraživanju ljudske inteligencije i unatoč vanjskoj sličnosti postoji temeljna razlika između prirodne i simulirane umjetne inteligencije, a to znači da se kompjutoru ne može u pravom smislu pripisati duhovna svojstva i stanja kao što ni ljudski

[2] David L.Poole, Alan K.Mackworth, *ArtificialIntelligenceFoundationsofComputationalAgents*, CambridgeUniversity Press,New York, 2010., str. 3,4.

mozak ne može u bilo kojem smislu biti kompjutor. Suprotno tome prema „jakoj“ teoriji kompjutori mogu biti inteligentni u doslovnom smislu riječi.[3]

1.1. Razine Umjetne inteligencije

Autori Cassio Pennachin, Ben Goertzel i Nick Bostrom ukazuju na tri moguće razine umjetne inteligencije, a to su: uska umjetna inteligencija (*Atificial Narrow Intelligence*), opća umjetna inteligencija (*Artificial General Intelligence*)i umjetna super–inteligencija (*Artificial Super-intelligence).*

Uska umjetna inteligencija može rješavati probleme unutar jedne ograničene sfere djelovanja. Primjerice, uska umjetna inteligencija može izvrsno igrati šah, ili prikupljati određenu vrstu podataka, upravljati osobnim automobilom, ali ne može ni na koji način djelovati izvan uskog područja za koje je osmišljena.

Opća umjetna inteligencija još nije izumljena niti je sasvim jasno kako uspješno stvoriti ovakvu umjetnu inteligenciju pa zato autori kao Cassio Penannachin i Ben Goertzel upozoravaju da se o njoj može govoriti samo kao o „hipotetičkom postignuću“. Opća umjetna inteligencija imala bi sposobnost planiranja, razumijevanja kompleksnih ideja, apstraktnog mišljenja, brzog učenja i učenja iz iskustva. U suštini posjeduje svojstva opće inteligencije potrebne za rješavanje širokog spektra problema kakva je nastala kroz evoluciju živih bića.[4]

Umjetna super–inteligencija naziv je za opću inteligenciju za koju neki autori predviđaju da bi po razini svoje inteligencije nadmašila čovjekovu. Smatra se da će umjetna super-inteligencija biti izumljena uskoro - nakon izuma opće umjetne inteligencije.[5]

Za sada čovjek je uspio napraviti jedino usku umjetnu inteligenciju i ona se koristi u mnogim aspektima ljudskog života. Ovakve sustave koristimo u svojim računalima, mobitelima, autima i drugim napravama. Za primjere možemo uzeti Google pretraživač i Facebook. Ove stranice rabe velik broj uskih umjetnih inteligentnih sustavakoji koriste sofisticirane metode vrjednovanja razumijevanja korisničkih podataka kako bi odlučile što da pokažu svakom individualnom korisniku. Druge vrste uske umjetne inteligencije sustava već se nalaze u

[3]Usp. Rolf Erassme, *Der Mensch und die „Kuenstliche Intelligenz“. Eine Profilierung und kritische Bewertung der unterschiedlichen Grundauffassungen vom Standpunkt des gemässigten Realismus*, Philosophische Fakultät der Rheinisch-Westfälischen Technischen Hochschule, (disertacija), Aachen, 2002. str. 4.
Pristupljeno 23.4. 2019: http://webdoc.sub.gwdg.de/ebook/ra/2004/erassme/02_194.pdf
[4]Cassio Pennachin, Ben Goertzel (eds.), *Artificial General Intelligence With 42 Figuresand 16 Tables*, Springer-Verlag Berlin Heidelberg, New York, 2007., str. 4.
[5]Nick Bostrom, *Superintelligence Paths, Dangers, Strategies*, Oxford University Press, 2014., str. 22.

samostalnom upravljaju zrakoplova u tetu, a postupno ih se uvodi i u cestovna vozila. Još sofisticiranije uske umjetne inteligencije koriste se u industriji, edukaciji, medicini, ekonomiji, vojsci, itd.,te se uvelike oslanjaju na njih u proizvodnji kao i pri filtraciji podataka te generalnom poboljšanju proizvoda i davanju usluga.

Ovakvu umjetnu inteligenciju relativno je lako programirati i proizvesti, no ona ima svoja ograničenja. Kao što je navedeno, veliko ograničenje uske umjetne inteligencije leži u tome što, iako jeu stanju lako provesti komplicirane matematiče izračune u sekundi, njezina sposobnost rješavanja problema ograničena je uskim poljem za koje je programirana. Tako uska umjetna inteligencija koja precizno upravlja bespilotnom letjelicom ne može, primjerice, odlučiti koje su dionice najprofitabilnije ili sačiniti metaanalizu znanstvenih radova na polju umjetne inteligencije. Nedostaje joj sposobnost shvaćanja svijeta izvan područja uske specijalizacije, kreativnog razmišljanja, učenja te povezivanja informacija iz različitih konteksta. Dakle, nedostaje joj upravo ona vrsta opće inteligencije koju posjeduje ljudsko biće.

U temelju istraživanja razvoja umjetne inteligencije danas leži konekcionistički model kognicije blizak rezultatima neuroznanosti. Ovaj model „analogiju pravi ne više s centralnim procesorom uz dodatak memorije, nego s horizontalno distribuiranom mrežom relativno nezavisnih jedinica koje stupaju u međusobne interakcije na mnoštvo načina“.[6] Uzor za te sustave je ljudski mozak. Konekcionistički sustavi manje su programirani, a više trenirani. No, da bi se stvorila umjetna inteligentna mreža koja bi sličila ljudskom mozgu trebao bi se shvatiti ljudski mozak, a mnogi neuroznanstvenici ističu da je on toliko složen da nam takvo znanje nedostaje.

Naši mozgovi posjeduju nevjerojatno kompleksne sustave analiziranja podataka u različitim kontekstima, kao i kreativnog rješavanja problema jer je prirodnom selekcijom takav pristup odabran kao najkorisniji za preživljavanje. Zbog toga je izuzetno teško pronaći programska rješenja koja će premostiti ovaj jaz između uske i opće inteligencije. No, iako su naši mozgovi prilagođeni obavljanju različitih svakodnevnih aktivnosti, postoje autori koji misle kako mi ipak nismo sposobni mjeriti se s uskom umjetnom inteligencijom koja se usavršila u svom polju, gdje probleme rješava brzinom od nekoliko milijuna matematičkih operacija u sekundi, stoga ona ostaje vrlo učinkovit alat koji će nam, bez sumnje, biti nezamjenjiv na putu do

[6] Olga Nikolić, Utjelovljena svijest i naturalizirana fenomenologija, *Filozofska istraživanja*, (3/2017), str. 547., bilj. 3.

stvaranja općeumjetne inteligencije. Naravno, ovdje bi se moglo postaviti pitanje o rješavanju kojih problema je riječ?

Opća umjetna inteligencija još nije izumljena, ali se radi na tome da se proizvede „misleći stroj“ koji bi imao opću primjenu u ljudskom životu. Razina inteligencije ovih sustava ne mora nužno dosegnuti inteligenciju čovjeka, dovoljno je da je sposobna rješavati široku lepezu problema u svim područjima života, odnosno da nije ograničena na jedno usko područje kao uska umjetna inteligencija. Razlog zašto ovakva umjetna inteligencija još nije izumljena leži u tome što računalna znanost još nije ovladala potrebnim tehnikama za stvaranje opće inteligencije te u ovom trenutku još uvijek radimo na usavršavanju uske umjetne inteligencije. Isto tako, za određene modele opće umjetne inteligencije, trenutne mogućnosti i cijene računalnog hardvera predstavljaju glavnu prepreku daljnjem napretku.

Onog trenutka kada načinimo opću umjetnu inteligenciju, morat će se imati na umu da će sustav, ako ikad dosegne razinu ljudske inteligencije, imati značajnu prednosti nad samim čovjekom. Prva prednost je brzina procesiranja informacija, kao i kapacitet za pohranjivanje novih informacija. Također, takvi sustavi bit će uvelike precizniji od organskih bića koja u svojim postupcima imaju veliku razinu pogreške. Nadalje, opća umjetna inteligencija moći će se međusobno povezati i razmjenjivati velike količine informacija, a da pri tome nemaju problema sa suradnjom kao što ima čovjek. Uz to, možda najvažnija karakteristika umjetne opće inteligencije je sposobnost samopoboljšanja, odnosno, ono nužno posjeduje sposobnost potpuno autonomnog nadograđivanja i unaprjeđivanja vlastitog sustava, što će posljedično dovoditi do sve veće razine inteligencije.

1.2. Metode stvaranja opće umjetne inteligencije

Trenutno nema izravnog odgovora na ovo pitanje, no postoji veliki broj različitih prirodoznanstvenih metoda razvoja UI čiji je cilj stvaranje opće inteligencije, od kojih su najistaknutije metode plagiranje mozga, evolucija i prepuštanje pronalaska rješenja računalu.

1.2.1.Metoda *plagiranja mozga*

Ova metoda nastoji postići svojevrsni obrnuti inženjering (*reverse engineering*) mozga kako bi se moglo proučiti na koji način je evolucija načinila takav organ. Kada bismo saznali kako mozak funkcionira na neurološkoj razini, imali bismo model za izradu opće inteligentnih

sustava. Jedan primjer takvog računalnog modela sadržavao bi mrežu tranzistora, odnosno „neurona“ međusobno povezanih *inputom* i *outputom*. Na početku,ovakva opća umjetna inteligencija ne bi imala nikakvo znanje, baš kao i novorođenče, no način na koji bi učila bio bi sličan čovjekovom. Ona bi pokušala izvršiti određeni zadatak te bi isprva neurološka aktivnost i pokušaji rješavanja zadatka bili sasvim nasumični, no ako pronađe ispravno rješenje, zaprimit će pozitivnu potvrdu pa će povezanost tranzistora koji su uspjeli proizvesti točan odgovor postati čvršća. Nasuprot tome, kada bi joj bilo rečeno da je pogriješila, veza između tih tranzistora bio slabila. Nakon nekog vremena i puno pokušaja i pogrešaka, mreža će sama od sebe formirati „živčane“ puteve i stroj će postati efikasniji u rješavanju zadatka. Naš mozak uči na sličan, ali sofisticiraniji način i kako se nastavlja proučavati mozak, otkrivaju se novi načini kako iskoristiti i plagirati mozak.[7]

Također postoji još jedan način plagiranja mozga, no ovo rješenje je manje vjerojatno zbog izuzetno velike razine kompleksnosti. Ova je strategija izravnija nego prethodna, zove se cjelovita emulacija mozga (*whole brain emulation*) i zahtijevala bi da se pravi mozak odvoji od osobe i izreže na tanke slojeve kako bi se sami dijelovi i stanice detaljno skenirali te bi se uz pomoć softvera rekonstruirao 3D model uzetog mozga. Taj model bi ili emulirao određene ili sve moždane funkcije i mehanizme te bi se te informacije implementiralo u moćno računalo koje bi bilo sposobno raditi određene stvari ili sve što i ljudski mozak može, samo bi moralo nastaviti učiti i skupljati informacije.[8]

Kako bi slikovitije prikazali metodu, Anders Sandberg i Nick Bostrom teoretiziraju da kada bi se radila cjelovita emulacija mozgakoji je pripadao preminuloj Mariji, računalo bi se teoretski moglo probuditi kao ta Marija u obliku opće inteligentnog računalnog sustava te bi se moglo raditi na tome da se Mariju pretvori u nevjerojatno briljantnu super–inteligenciju. Usprkos tome, cjelovita emulacija mozga neće biti sasvim vjerna prirodi, zato što nije uvijek cilj proizvesti računalo koje bi do najsitnijih detalja emuliralo Marijin skenirani mozak. Vrsta emulacije ovisi o onome što želimo postići– potpuno biološko i psihološko razumijevanje mozga ili opće inteligentno računalo kojem, ustvari, ne treba osobnost niti sjećanja Marijinog mozga. Stoga će, pretpostavlja se, emulacija mozga biti nekakav hibrid između cjelovite biološke emulacije i simulacijeosnovnih moždanih funkcija.

[7]N. Bostrom, *Superintelligence Paths, Dangers, Strategies*, 2014. str. 8.
[8]Anders Sandberg, Nick Bostrom, *Whole Brain Emulation: A Roadmap, Technical Report*, 2008-3, Technical Report 2008-3, Future of Humanity Institute, Oxford University 2008., str. 16.

1.2.2. Emulacija evolucije

Drugi način postizanja opće umjetne inteligencije jest emulacijom evolucijskih procesa koji se odvijaju u organskim bićima. Ako se ispostavi da je plagiranje ili emuliranje ljudskog mozga prekompleksno ili se ispostavi kao loša strategija, spekulira se da bi se onda moglo probati s evolucijskom komputacijom (*evolutionary computation*)[9] i uz pomoć genetičkih algoritama (*genetical algorithm*) Jenna Carr nam daje pojašnjenje ovih algoritama:

> „Genetički algoritmi su tip optimizacijskog algoritma, što znači da se koriste kako bi pronašli optimalno rješenje ili rješenja određenom komputacijskom problemu koji maksimizira ili minimalizira određenu funkciju. Genetički algoritmi reprezentiraju jednu od grana znanstvenog polja zvanog evolutionary computation, u smislu da imitiraju biološki proces reprodukcije i prirodne selekcije kako bi da li „najjača" rješenja. Kao i u evoluciji, mnogi od genetičkih algoritamskih procesa su nasumični, no ova optimizacijska tehnika dopušta da se odrede stupanj nasumičnosti i razinu kontrole. Ovi algoritmi su daleko moćniji i učinkovitiji od nasumičnog traženja i iscrpnih pretražujućih algoritama te ne zahtijevaju dodatne informacije o zadanom problemu. Ovo svojstvo im omogućuje da pronađu rješenja na probleme s kojima se koje ostale optimizacijske metode ne mogu nositi zbog nedostatka kontinuiteta, derivata, linearnosti ili drugih svojstava."[10]

U knjizi „*The Artificial Life Route To Artificial Intelligence Building Embodied, Situated Agents",*strategija emulacije selektivne evolucije objašnjenjena primjeru bihevioralnog računalnog sistema za izbjegavanje prepreka:

> „Polemiziramo da postoje mehanizmi koji generiraju nove kopije bihevioralnih sustava unutar agenta. Kopije bi mogle imati manje varijacije poput nešto drukčijih postavki parametara, transformacije struktura, dodatne senzore, itd. Budući da imamo aditivnu nehijerarhijsku kombinaciju bihevioralnih sustava koji će djelovati kad god se uvjeti u okolišu podudaraju sa zahtjevima njihovih senzora. Bihevioralni sistem kojeg se efektivno koristi postaje jači. Snaga znači i preživljavanje u bazenu nadmetajućih bihevioralnih sistema i veću vjerojatnost da nastane novi potomak.

[9]Skupina algoritama za globalnu optimizaciju inspirirani biološkom evolucijom, koja pronalazi rješenja na temelju pokušaja i pogrešaka.
[10]Jenna Carr, *An Introduction to Genetic Algorithms,* Abstract, Senior Project, 2014., str. 1. Pristupljeno: 7.4. 2019: http://www.joinville.udesc.br/portal/professores/parpinelli/materiais/IntroductionGA.pdf

> Primijetite da ne postoji vanjski evaluacijski kriterij koji daje ocjene jednom ili drugom sistemu kako je inače slučaj s uporabom genetičkih algoritama.“[11]

Strategija emulacije evolucije uz korištenje genetičkih algoritama bi, dakle, trebala funkcionirati na sljedeći način-genetičkim algoritmima bi se evaluirale performanse grupe računalnih sustava, slično kako priroda evaluira „performans“ živih bića i „procjenjuje“ hoće li biće moći prenijeti dalje svoje gene. Prema toj teoriji, grupa računala koja bi pokušavala izvršiti neki zadatak i ona koja budu najučinkovitija u rješavanju bi trebala sjediniti svoje programe ili ih kopirati (s manjim varijacijama) u jedan novi sustav. Manje uspješni sustavi bi bili eliminirani. Nakon što se ovaj proces ponovi puno puta, „prirodni“ odabir bi uz pomoć genetičkih algoritama davao sve bolje sustave. Genetički algoritmi bi ujedno trebali biti i ono što će pomoći pri uklanjanju neželjenih i množenju željenih grešaka i promjena unutar računalnih sistema.

Također, naspram prirodnog evolucijskog procesa koji je nasumičan i producira beskorisne ostatke evolucije, mi bismo prema ovoj strategiji vodili emuliranu evoluciju na takav način da bi se proizvelo onakvo računalo kakvo želimo. Također bismo imali mogućnost da sami interveniramo i uklonimo beskorisne evolucijske ostatke koji bi se mogli nalaziti u programima.

No problematično je izraditi automatizirani evolucijski proces kao i sjedinjavanje programa i način kako bi ovaj cjelokupni ciklus mogao samostalno funkcionirati. Uz to, treba pronaći adekvatne parametre, tj. algoritme kako bi cijeli evolucijski proces, koji je inače nasumičan i traje milijunima godina u prirodi, trajao kratko (desetak godina) i dao željene rezultate. Iako ova metoda zvuči plauzibilno, nije sigurno hoće li se moći poboljšati evolucijski proces dovoljno da ovo postane valjana strategija.

1.2.3.Prepuštanje pronalaska rješenja računalu

Ova strategija je najjednostavnija i ima najveće šanse za uspjeh za razliku od prethodnih. Prema ovoj strategiji načinilo bi se računalo čije će dvije glavne vještine biti istraživanje sustava umjetne inteligencije i rađenje promjena vlastitog koda kako bi se učilo i poboljšalo vlastiti sustav. Dakle, glavni zadatak takvog računala bio bi pronalazak načina kako samog sebe učiniti pametnijim.

[11] Luc Steels Luc, Rodney Brooks (ed.), *The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents,* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1995., str. 106.

Iz ovih informacija se lakše uviđa i ponajviše naglašava kompleksnost opće i super-inteligentnih umjetnih sustava što i ukazuje na njihove potencijalne sposobnosti. Valja voditi računa o tome što je uopće potaknulo cijelu debatu oko toga može li stroj imati svijest i time moralni status, a to je upravo njihova kompleksnost i inteligencija koje su važne u etičkom razmatranju moralnih statusa sustava umjetne inteligencije.

2. UVOD U ETIČKO RAZMATRANJE I PITANJE SVIJESTI UMJETNO INTELIGENTNIH SUSTAVA

Kada se postavi pitanje moralnog odnosa prema umjetnoj inteligenciji, zapravo se postavlja pitanje o uvjetima po kojima bi se umjetna inteligencija mogla smatrati moralnim pacijentom, odnosno bićem koji može trpjeti posljedice ljudskih odluka. Sintagmom „UI sustav“ referira se uglavnom na opće umjetno inteligentne i super-inteligentne umjetne sustave jer se na njihovoj osnovi razvija rasprava o moralnom statusu umjetno inteligentnog sustava, ali i na tezi da se uska umjetna inteligencijamože shvatiti i tretirati kao alat.

U ovom radu seneće tematizirati pitanje može li se umjetnu inteligenciju smatrati moralnim djelateljem iz nekoliko razloga. Prvi je taj što je takva rasprava više rezervirana za budućnost, kada i ako se UI tehnologija dovoljno razvije da bi se moglo adekvatno raspravljati o njezinom statusu djelatelja. Trenutno se pitamo ima li uopće smisla umjetni sustav promatrati kao nešto što može uopće trpjeti posljedice odluka na način koji živi sustavi trpe upravo zato jer je ova tehnologija tek u začetku. Možda će se tehnologija budućnosti kretati u tom smjeru da će biti potrebo izgraditi etički umjetno inteligentni sustav, što će onda zahtijevati rasprave oko mogućeg statusa moralnog djelatelja, stoga se ta rasprava ostavlja po strani jer je cilj istražiti problem zašto uopće možemo i trebamo otvoriti raspravu o umjetnoj inteligenciji kao o moralnom pacijentu. Ovo pitanje je važno jer generalno možemo reći da se želimo moralno odnositi prema bićima koja mogu trpjeti posljedice naših odluka i želimo znati koja su to bića koja možemo staviti u područje etičkog promišljanja, odnosa i zašto.

Rasprava se temelji na onome što nam utilitaristička i deontološna etička teorija mogu reći o ovom problemu te će se usporediti s raspravom o životinjskim pravima jer se tako može doći do zaključka o moralnom statusu umjetne inteligencije.

Prije analize dviju navedenih teorija, njihovih prednosti i nedostataka, važno je jasno postaviti problem i argumente koji ga okružuju.

Smatra se da prvi argument koji se nameće pri utvrđivanju smislenosti debate oko moralnog statusa umjetnog inteligentnog sustava je u suštini vrlo jednostavan- umjetno inteligentni sustav treba se promatrati kao alat ili kao mehaničkog slugu. Umjetno inteligentni sustavi nisu ništa više nego spoj svojih umjetnih dijelova i električnih spojeva, stoga je apsurdno smatrati nešto kompleksniji alat moralnim pacijentom.

Ovaj stav nabolje je sažela Joanna Bryson čiji će se rad uzeti kao primjer navedenog argumenta. Bryson tvrdi da je umjetno inteligentni sustav jednostavno samo skup podataka i umjetnih materijala koji rade točno što im je zadano, kao i druga tehnologija koju je napravio čovjek te zato tvrdi:

> „Zapamtite, robot je upotpunosti posjedovan i dizajniran od strane čovjeka. Mi određujemo njegove ciljeve i želje. Robot ne može biti frustriran osim ako mu nije zadan cilj kojeg ne može ispuniti i ne može mu smetati biti frustriran osim ako ga ne programirao da percipira na frustraciju kao nešto loše,umjesto kao indikaciju problema u planiranju. Robota se može zlostavljati baš kao što i auto, klavir ili kauč mogu biti zlostavljani – može ga se oštetiti na nesvrsishodan način. No opet, nema razloga da ga se programira da mu smeta takav tretman.“[12]

Nadalje, Bryson opravdava ovakav pogled na umjetne sustave s argumentom da su oni, naposljetku, samo čovjekovi alati:„Roboti su alati i kao svaki drugi objekt kada je u pitanju domena etike. Možemo koristiti ove alate kao produžetak naših sposobnosti i povećati našu učinkovitost analogno kako je i velika proporcija profesionalnog društva kroz povijest povećavala svoje vlastite sposobnosti slugama.“[13] Navedenu tezu potvrdili su i Deborah Johnson i Keith Miller: „Tvrdnje su čudne jer se čini da ne priznaju da je računalni sustav produžetak ljudske aktivnosti i zato jer njihov pogled na svjesno ljudsko djelovanje [agency]i moralnost nije u skladu s idejom moralnosti kao ljudskog sustava kontekstualiziranih ideja i značenja.“[14]

Primjedba iznesena u prethodnim citatima svodi se na tvrdnju kojom da bi se etičke rasprave trebale zadržati samo na živim bićima i da je besmisleno voditi takve rasprave oko umjetnih sustava. Naravno, usku umjetnu inteligenciju se može smatrati alatom ili produžetkom ljudskog djelovanja[15], no to ne znači da se na isti način može promatrati i kompleksni sustav kao što je umjetna opća i super-inteligencija. Smatra se, ukoliko postoji osnovana sumnja, da bi umjetni sustav mogao imati svijest,a tase sumnja *mora* razmotriti te potvrditi ili opovrgnuti.

---

[12] Joanna J.Bryson, Robots Should Be Slaves, *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, John Benjamins, 2010., str. 63-74. Ovdje str. 9
Pristupljeno 4.9.2019:https://www.researchgate.net/publication/250333956_Robots_Should_Be_Slaves

[13]Ibid., str 10.

[14]Deborah G.Johnson, Keith William Miller, Un-makingartificial moral agents, *Ethics and Information Technology, Vol. 10*., 2008., Springer Netherlands str. 123-133. Ovdje str. 127.
Pristupljeno 9.4.2019: https://link.springer.com/article/10.1007/s10676-008-9174-6

[15]No kako napreduje razvoj tehnologije smatram da ovakvo jednostavno viđenje umjetne inteligencije postaje pomalo redukcionističko.

Etika nebi trebala biti ograničena u svom razmatranju moralnih statusa raznih bića te se moramo podsjetiti da se njezina domena, kroz povijest, uvijek proširivala. Kako je već prije rečeno, krug etičkog razmatranja se širio od samoga čovjekapa do same prirode koja nas okružuje, upravo jer se vodimo željom da etički tretiramo sva ona bića koja mogu, na neki način, osjetiti posljedice našeg djelovanja.

Vraćajući se na tvrdnje Joanne Bryson, vidjeli smo da jenavedeno mišljenje da se prema umjetnoj inteligenciji može neetički odnositi kao i kad kažemo da se neetički odnosimo prema, primjerice biciklu. Ako oštetimo bicikl možemo reći da je to neetičko djelovanje jer smo *vlasnika* objekta novčano oštetili ili da je neetički uništavati nešto što ima ekonomsku ili estetsku vrijednost. Dakle, prema *nečemu* ćemo se moralno odnositi jer bi bilo čovjeku u interesu zaštiti ono što mu je korisno na direktan ili indirektan načinpa bismo stoga željeli zaštiti i umjetne sustave kao objekte jer su nam korisni te bi tim zaključkom etička rasprava trebala bila završena.

Etičko se promišljanje statusa umjetne inteligencije ponekad svodi na navedeni argument koji nas upućuje na zaključak da ustvari ne postoji etička dužnost prema samom sustavu kao biću[16], nego prema vlasniku i onima koji barataju takvom tehnologijom. No, intuicijski smo svjesni da ovaj problem, zbog sve veće kompleksnosti i inteligencije ovih sustava, nije tako jednostavan kako se doima na prvi pogled. Inteligencija umjetnih sustava i njihova složenost nas upućuju na to da bi ih se trebalo razmotriti kao bića koja mogu trpjeti posljedice naših odluka, slično kako to mogu drugi živi sustavi. Ako pokažemo da je umjetni sustav nešto više od alata (specifično umjetna opća inteligencija i umjetna super-inteligencija), onda možemo odbaciti prethodno navedeni etički odnos kao neetičan jer je takav etički odnos temeljen isključivo na ljudskom interesu, a ne na samom subjektu, tj. umjetno inteligetnom sustavu.

U drugu ruku, sama činjenica da je to umjetni sustav koji slijedi zadane naredbe ili algoritme upućuje nas na drugi zaključak- umjetno inteligentni sustav *jest* samo predmet kojim se čovjek služi te ga tako valja i tretirati.

Očigledno je da umjetno inteligentnim sustavima nešto nedostaje da mogu ući u krug etičkog razmatranja. Dakle, koji su uvjeti koje bi umjetna inteligencija morala ispuniti da bi ju se moglo smatrati moralnim pacijentom?

[16]U nastavku teksta ću se ponekad referirati na umjetno inteligentni sustav kao na biće, no ne u smislu živog bića, nego na nešto što bivstvuje, postoji i sudjeluje u svijetu na svojstven način.

Iako se ne možemo složiti oko toga što točno umjetnoj inteligencjidaje nekakav moralni status, najčešće, kako Nick Bostrom[17]jednostavno navodi, „bića moraju imati razum ili svijest (samosvijest, iskustvo samoga sebe, sposobnost reagiranja na okoliš i sl.) ili mogućnost da iskuse patnju.“[18]

Možemo primijetiti da je čak problematično i tvrditi da jedino bića koja ispunjavaju ova dva uvjeta nužno imaju moralni status pacijenta. No, da ne skrenemo s teme, svejedno ćemo uzeti da su ovi uvjeti potrebni za etičko promišljanje umjetno inteligentnog sustava jer se rasprave oko moralnog statusa umjetne inteligencije kreću upravo oko navedenih kriterija.

Dakle, prema ovim kriterijima umjetno inteligentni sustav mora imati iskustvo svijesti i/ili sposobnost da iskusi patnju čime bi ga se dovelo na razinu moralnog statusa, npr. životinja. Stoga zaključujemo da je irelevantno odakle i kako u biću nastaje fenomen svijesti, bitno je samo da on postoji. Također je i irelevantno pitanje radi li se o organskom ili anorganskom biću. Uz ovo, stoji i da ako biće ima svijest i iskustvo patnje, nevažno je kako je to biće nastalo, umjetnim ili prirodnim putem. U slučaju da se ipak ne slažemo s ovim kriterijima te tvrdimo da su pogrešni ili nebitni, doći ćemo do problema opravdanja moralnog statusa životinja, kloniranih ljudi ili sličnih posebnih slučajeva.[19]

Zaključili smo da su uvjeti potrebni za ostvarenje moralnog statusa svijest i/ili sposobnost osjećanja patnje. Ova će se tvrdnja svesti na to da je u suštini najvažniji uvjet koji umjetno inteligentni sustavi moraju ispuniti jest taj da u njima postoji svijest.

Kako bi se moglo replicirati na Brysonov argument, morat će se pokazati da u naprednim umjetno inteligentnim sustavima možemo pronaći nešto kao što je svijest ili možda čak i osjećaj boli. Zato će se u sljedećim poglavljima najviše tematizirati problematika svijesti u umjetnim sustavima te će se iznijeti problematika *kineske sobe,* kao i nekoliko teorija prema kojima se svijest može razviti i u umjetnim sustavima kako bi se pokazalo da ima smisla raspravljati o mogućem moralnom statusu umjetne inteligencije. Dotaknut ćemo se problematike osjećajnosti umjetnih sustava, tj. onog što bi moglo imati iskustvo boli, no na njemu neće biti naglasak, jer se smatra da tehnologija neće ići u takvom smjeru, odnosno nije

[17]Eliezer Yudkowsky, Nick Bostrom, *The Ethics of Artificial Intelligence*, Draft for Cambridge Handbook of Artificial Intelligence, eds. William Ramsey and Keith Frankish Cambridge University Press. 2011., str. 7 Pristupljeno 9.4.2019: https://nickbostrom.com/ethics/artificial-intelligence.pdf

[18] Složit ćemo se da današnja uska umjetna inteligencija nema ove sposobnosti, no kao što je prije navedeno u radu, smarta se da buduća UI tehnologija pa tako umjetna opća inteligencija i umjetna super-inteligencija mogu imati ove atribute. UI sustavi već sada mogu reagirati i vršiti interakciju s okolišem, no kroz rad će se pokušati pokazati da nekakav tip svijesti može postojati u UI sustavu.

[19] Ibid., str. 8.

održivo da želimo imati umjetne sustavekoji mogu osjetiti neugodu na način na koji ju živa bića osjećaju. Osim toga,ovakve sustave bilo bi vrlo teško načiniti, zato što ne znamo što patnja jest kao psihološki fenomen ili kako analizirati takav fenomen, kao što ne znamo način na koji bismo to znanje mogli iskoristiti pri izradi 'osjećajućeg' umjetnog sustava.

3. PITANJE SVIJESTI UMJETNE INTELIGENCIJE

Prije nego što se posvetimo etičkom razmatranju umjetno inteligentnih sustava, nameće se pitanje svijesti jer ako samaumjetna inteligencija nije svjesna,onda ne postoji etički problem. Rasprava o pitanju svijesti započinje izlaganjem komputacijske teorije uma StevenaHorsta kojom sena svijest gleda kao naodređene komputacijske sposobnostiuma. Nakon toga će se izložiti replika Johna Searlea na komputacijsku teoriju uma. Potom ćemo se osvrnuti na problem svijesti s određenih filozofskihi znanstvenih stajališta kao svojevrsni odgovor Searleu, a koje se smatraju relevantnima za ovaj rad kako bi se pokušala pokazati ideja da se fenomen svijest pojavi u umjetnom sustavu nije nemoguć i time zadržati smislenost rasprave o moralnom statusu umjetne opće i super-inteligencije.

Što je, dakle, svijest? Ovdje ne pretendiramo dati odgovor jer je tema svijesti predmet višestoljetne rasprave i njezino raspravljanje bi uveliko premašilo okvir i svrhu naše teme. Na to bi se moglo primijeniti ono što je Augustin rekao o vremenu. Kada nas ne pita, mislimo da znamo što je svijest, a kada nas se pita ne znamo odgovor. Općenito možemo razlikovati pristupe koji svijest opisuju iz perspektive prve osobe i one koji ju opisuju iz perspektive treće osobe. David Chalmers u knjizi *The Conscious Mind* uvodi razliku između „teškog i lakih probema svijesti". Problem svijesti spada u „teški problem", a Dan Zahavi interpetirajući Chalmersa pojašnjava da se ovaj problem odnosi na „problem objašnjavanja zašto mentalna stanja imaju fenomenalna ili iskustvena svojstva".[20] Kada se govori o svijesti , najčešće se ne govori samo o njoj, nego i samosvijesti. No, o njoj je „važno govoriti jedino s obzirom na bića koja se kao individue koje kažu 'ja' same od sebe izričito odnose na sebe same". [21] U tome V. Gerhardt vidi „jednostavan kriterij za pripisivanje samosvijesti u sveukupnosti živih bića" [22]

3.1. Komputacijska teorija uma

Komputacionizam je bio dominantni model sredinom dvadesetog stoljeća u kognitivnim znanostima i analitičkoj filozofiji. Um se „shvaćao prema analogiji sa centralnim procesoromi memorijom, odnosno po modelu *input* – obrada – *output*".[23] Komputacijska

[20]Dan Zahavi, Intencionalnost i iskustvo, *Filozofska istraživanja,* 2/ 2006.
Pristupljeno 16.4.2019:
https://hrcak.srce.hr/search/?show=results&stype=1&c%5B0%5D=article_search&t%5B0%5D=Dan+Zahavi%2C+Intencionalnost+i+iskustvo
[21]Volker Gerhardt, *Samoodređenje: princip individualnosti*, Demetra, Zagreb, 2003., str. 144.
[22]Ibid., str. 144.
[23]O. Nikolić, Utjelovljena svijest i naturalizirana fenomenologija, *Filozofska istraživanja*, (3/2017), str. 546.

teorija uma nam govori da je ljudski um digitalno računalo i da je misao doslovno vrsta algoritma po kojem se izvode komputacije nad simbolima. Ako je vjerovati ovoj teoriji, nije apsurdno tvrditi da i umjetna inteligencija kao digitalno računalo ima svijest. Kako bi pobliže razumjeli točno što je komputacijska teorija uma, Steven Horst nam daje objašnjenje ove teorije u tekstu „*The Computational Theory of Mind*".

> Prema njemu, navedena teorija u sebi sadrži *Representational Theory of Mind*, odnosno: „…tezu da stanja poput vjerovanja i željenja su relacija između mislitelja i simboličkih reprezentacija sadržaja tih stanja: na primjer, vjerovati da postoji mačka na otiraču jest biti u određenoj relaciji (karakterističnoj za vjerovanje) prema simboličkoj mentalnoj reprezentaciji čija je semantička vrijednost 'postoji mačka na otiraču', nadati se da postoji mačka na otiraču znači biti u drukčijoj funkcionalnoj relaciji (karakterističnu onoj za nadanje, nego onoj vjerovanja) prema simboličkoj mentalnoj reprezentaciji s istom semantičkom vrijednosti."[24]

Uz ovu tezu još se dodaje i teza o *Computational Account of Reasoning* koja je povezana s prethodnom tezom, a ona sama tvrdi da:

> „…ove reprezentacije imaju semantička i sintaktička svojstva i procesi rasuđivanja se provode tako da reagiraju samo na sintaksu simbola– tip procesiranja koji ima tehničku definiciju 'komputacije' i poznata je kao formalna manipulacija simbola (npr. manipulacija simbola prema čisto formalno – npr. nesemantičke – tehnike. Riječ 'formalno' modificira 'manipulaciju', ne 'simbol')."[25]

Dakle, um je simbolički operator i mentalne reprezentacije su ustvari simboličke reprezentacije objekata. Jednako kako se jezična semantika odnosi na značenja i definicije riječi, tako se i semantika mentalnih stanja odnosi na značenje tih reprezentacija. Prema tome, neka osnovna mentalna stanja mogu imati svoje određeno značenje, baš kao i zasebna riječ u jeziku. Ovo znači da mogu nastati i kompleksnija mentalna stanja poput misli. Kompleksnija stanja bi se mogla razumjeti ako postoji razumijevanje osnovnih komponenti koja su sintaktički točna.[26]

---

[24]Steven Horst, *The Computational Theory of Mind*, The Metaphysics Research Lab Center for the Study of Language and Information Stanford University, Stanford 2003. str. 2. Pristupljeno 4.9.2019:https://www.researchgate.net/publication/277224413_The_Computational_Theory_of_Mind
[25] Ibid., str. 2.
[26] Ibid., str. 5,6.

Zaključuje se da bi uz kognitivne sposobnosti ovakav stroj mogao posjedovati i svijest koja je također komputacijska. No, valja napomenuti da ova teorija ostavlja otvorenu mogućnost da neki aspekti uma nisu komputacijski.

Nadalje, prema ovoj teoriji uma, kognitivni mehanizmi (npr. računanje, percepcija, učenje jezika) su usađeni u mozgu, tj. fizički manifestirani kroz aktivnost neurona. Naš mozak i um su poput univerzalnog Turingovog stroja koji manipulira simbolima prema određenim pravilima, u kombinaciji s unutarnjim stanjima stroja. Ono što je bitno za ovu teoriju jest to da se možemo odmaknuti od određenih fizičkih karakteristika takvog stroja. Drugim riječima, komputacija može biti usađena, tj. fizički ostvarena kroz silikonske čipove ili neuronske mreže. Ono što je bitno jest da postoje outputi bazirani na manipulaciji inputa i unutarnjih stanja i izvedeni na osnovi određenih pravila. Ako za navedene mehanizme zapravo nije bitna fizička manifestacija, onda bi mogli biti usađeni i u nekoj drugoj vrsti komputacijskog stroja poput računala.[27]

Umu su potrebne mentalne reprezentacije objekata, tj. simboli objekata koji postaju *input* jer komputacija ne može biti izvršena nad stvarnim objektom, nego se objekt mora interpretirati i postaviti u oblik s kojim može raditi, tj. manipulirati. Uz pomoć semantike, tj. značenja tih reprezentacija, moguće je da um procesira složene reprezentacije i simbole. Iz navedenog se može zaključiti da se rad mozga može izraziti algoritmom, zato što je on sam procesor simbola.

Iz navedenog se zaključuje da ako računalo može funkcionirati na sličan način kao ljudski um, onda ljudski um može funkcionirati donekle poput računala. Dakle, ljudi susvjesna bića koja imaju iskustvo i razumijevanje svijeta, stoga i umjetno inteligentni sustav može imati ili razviti sve ili neke ljudske karakteristike. Prema ovome, umjetnu inteligenciju bi se trebalo tretirati, u najmanju ruku, kao moralnog pacijenta. No, kako možemo znati da umjetni sustav doista ima svijest kao što ju imaju živa bića te da ju umjetni sustav naprosto ne simulira?

No, glavni prigovor ovakovij teoriji sastoji se u tome da ona zanemaruje „konstitutivnu ulogu živog tijela za kogniciju i svijest“.[28]

[27]Ibid., str. 10, 11.
[28]O. Nikolić, Utjelovljena svijest i naturalizirana fenomenologija, str. 547.

3.2. Searleov misaoni eksperiment i replike

Najpoznatija i najpopularnija kritika komputacijskog funkcionalizma i implikacija ove teorije jest kritika Johna Searlea i njegov misaoni eksperiment zvan *Kineska soba*. Ova je kritika jedna od mnogih, ali smatra se da daje vrlo dobre argumente protiv postojanja nekakve svijesti u umjetnom sustavu te se u bitnome svodi na problem kojeg Searle opisuje u svom tekstu „*Minds, brains and programs"*.

Searle na početku svog rada govori da mu je cilj istražiti propoziciju koja tvrdi da je čovjekova i životinjska intencionalnost produkt određenih moždanih procesa. On želi dokazati suprotnu tvrdnju, a to je da sam računalni program nikad neće biti dostatan uvjet za intencionalnost. Searle uz ovo želi pokazati i da svaki mehanizam koji je u stanju producirati intencionalnost mora imati procese ili moći jednake mozgu, zato bi svaki umjetno inteligentni sustav trebao biti identičan mozgu ako želimo s određenom sigurnošću tvrditi da ima intencionalnost. Dakle, on ne tvrdi da stroj ne može misliti, nego da jedino posebna vrsta stroja to može– mozak i stroj koji ima procese jednake onima koje vidimo u mozgu.[29]

Kako bi dokazao svoje tvrdnje, Searle nam daje primjer kineske sobe. Zamislimo usamljenog uredskog činovnika u zatvorenoj sobi koji kroz prorez na vratima dobije papir s pitanjima napisana kineskim simbolima koje uopće ne razumije. Međutim, unutar sobe naš činovnik raspolaže s pregršt instrukcija na jezikunkoji razumije (uzmimo hrvatski), a ove instrukcije mu ukazuju na koji način da odgovori na kineske simbole. Kako naš činovnik ne razumije kineske simbole, oni mu djeluju besmisleno i nije u stanju niti pretpostaviti što bi mogli značiti. No, slijedeći uputstvana hrvatskom, činovnik će biti u stanju kroz prorez na vratima dati nekakav odgovor napisan kineskim simbolima. Dakle, ovdje se instrukcije na hrvatskom trebaju shvatiti kao program koji je dan računalu. Nakon nekog vremena ljudi koji pišu instrukcije i činovnik u sobi koji ih slijedi toliko će se izvještiti u svom poslu da neće biti moguće razlikovati odgovore činovnika u sobi i osobe koja inače govori i razumije kineski. Nekome izvana će se činiti kao da razgovara s osobom koja razumije kineski, unatoč tome što to nije istina. Naš činovnik samo manipulira nerazumljivim mu formalnim simbolima. On bi se time ponašao poput računala, izvodeći komputacijske operacije na formalno specificiranim elementima. Dovoljni su samo adekvatni inputi (pitanja na kineskom) i programi (uputstva na

[29]John Rogers Searle, , Minds, brains, and programs,*The Behavioral and Brain Sciences 3,*Cambridge University Press, Berkeley California 1980., Str. 417-457.Ovdje str. 417.
Pristupljeno 4.9.2019:http://www.uh.edu/~garson/MindsBrainsandPrograms.pdf

hrvatskom) da se output (odgovori na kineskom) ne bi mogli razlikovati od odgovora koja bi dala osoba koja uistinu razumije kineski.[30]

Searle nam ovime poručuje kako nisu točne tvrdnje da snažni UI[31] razumije problem koji se treba riješiti ili da umjetna inteligencija u neku ruku može objasniti ljudsko razumijevanje. Odnosno, unatoč tome što računalo može dati tražena rješenja, ne znači da se iza toga krije razumijevanje onoga o čemu odgovara u smislu razumijevanja koje čovjek ima pa se zbog toga onda ne može doći do odgovora o funkcioniranju čovjekova uma.

Naš mozak, govori Searle, možda je stroj koji sadrži razne programe za baratanje raznim simbolima, no mi smo izloženi inputu kojem umjetna inteligencija nije, zato posjedujemo razumijevanje koje računalo nema. Zato on tvrdi da razumijevanje nečega ne ovisi o programu u smislu komputacijskih operacija.

Searl je dao primjer temeljem kojega se može napraviti jasna razlika između personalne inteligencije i njenog pukog simuliranja. „Jer čak i kada bi postojali savršeni kompjutorski prevoditelji, ne bi ni najmanje razumijeli jezik (također ni sintaksu), čije bi znakove mogli samo prema pravilima hantieren“.[32]

Postoji nekoliko replika na ovaj misaoni eksperiment i Searle sam daje odgovore na njih, ali izdvojit će se samo određene.

Prva replika govori da je ovaj misaoni eksperiment u suštini isti kao i problem tuđih svijesti. Dakle, može se postaviti pitanje kako možemo znati da čovjek u misaonom eksperimentu uopće razumije kineski? Dolazi se do toga da ako čovjeku pripisujemo kogniciju i moć razumijevanja, onda nužno moramo i računalu. Searle odgovara da se mora pretpostaviti da u čovjeku postoje spoznatljiva kognitivna stanja, ponajprije jer raspravljamo o ljudskom umu[33] i jer komputacijski procesi i output mogu postojati i bez njih. Zato odgovara da: „…problem ove rasprave nije u tome kako znam da drugi ljudi imaju kognitivna stanja, nego u tome što je to čemu ih ja pripisujem kada im pripisujem kognitivna stanja.“[34] Odnosno čemu tom 'nečemu' se pripisuje intencija i razumijevanje.

---

[30] Ibid.,str. 418

[31] Searle razlikuje slabu UI od snažne UI koja bi se odnosila na umjetnu opću inteligenciju i umjetnu super-inteligenciju.

[32] Josef Seifert, *Das Leib-Seele Problem und die gegenwaertige philosophische Diskussion*, Wissenschaftliche Buchgesellschaft, Darmstadt, 1989., str. 28.

[33] Ako ne pretpostavimo njihovo postojanje ne možemo onda niti raspravljati o umu.

[34] J.R Searle*, Minds, brains and programs,* str. 421, 422

Dakle, ne možemo bilo čemu pripisati posjedovanje intencionalnost, posebice umjetnim sustavima jer, za razliku od čovjeka, već od prije znamo da tehnologija nema intencionalnost već da ona funkcionira samo po zadanim uputama bez razumijevanja, tj. znamo da u njoj ne postoje kognitivna stanja. Već smo naveli da komputacijski procesi i output mogu postojati bez kognitivnih stanja te, ako tim operacijama pripisujemo intencionalnost, onda možemo bilo čemu pripisati intencionalnost što je besmisleno. Zato se moramo čuvati od pripisivanja intencionalnosti umjetnim sustavima.

Zatim imamo idući protuargument koji govori da ako imamo robota čije je ponašanje posve isto kao i čovjekovo, s računalom za mozak koje je programirano sa svim moždanim sinapsama ljudskog mozga, sigurno bi mu se pripisalo svojstvo intencionalnosti. Searle se slaže s ovom postavkom. Morali bismo ovakvom robotu pripisati intencionalnost, ali pod uvjetom da ne znamo kako drukčije objasniti njegovo ponašanje osim intencionalnošću.

> „Zamislite da znamo da je robotovo ponašanje objašnjeno činjenicom da čovjek unutar njega prima neinterpretirane formalne simbole iz robotovih receptora i šalje neinterpretirane formalne simbole njegovim motornim mehanizmima i čovjek koji izvodi manipulaciju simbolima u skladu s hrpom pravila. Također, zamislite da taj čovjek ništa ne zna o ovim činjenicama o robotu, sve što zna jest koje operacije mora napraviti i na kojim besmislenim simbolima. U ovakvom slučaju gledali bismo na robota kao na genijalnu lutku. Hipoteza dalutka ima um bi tada postala neopravdana i nepotrebna jer više ne postoji razlog da pripišemo intencionalnost robotu ili sustavu….“[35]

Dakle, Searle smatra da su čovjek i životinje *jedina* bića koja imaju mentalna stanja i inpute potrebne za produkciju intencionalnosti.

No, i na ovu tvrdnju imamo još jednureplikukoju nam daje Arielle Zuckerberg u svom radu o umjetnoj inteligenciji. Zuckerberg uočava da kroz svoj tekst Saerle zapravo iskazuje vrstu straha zbog mogućnosti da je ljudski um samo neurološki stroj. Ona tvrdi da je kineska soba samo način kojim Searle izražava svoje iracionalno uvjerenje da je čovjek *ipak* posebno i jedinstveno biće u univerzumu. No, također i napominje da se standardi koje primjenjujemo na čovjeka ne mogu adekvatno primijeniti na umjetne sustave jer oni ne funkcioniraju sasvim isto kao ljudsko biće čineći njegov misaoni eksperiment ništavim. Također ga i optužuje da on zapravo ne razumije da su inteligencija i svijest odvojive te da njegov misaoni eksperiment

[35]Ibid., str. 421.

čak i brani validnost Turingova testa koji je inače kritiziran jerje veoma „bihevioristična“ u pogledu dokazivanja svijesti.[36]

Dakle, prema Searleu, stroj može razmišljati, imati svijest i intencionalnost- mozak je primjer jednog takvog biološkog stroja. No *umjetni* strojto ne može, osim ako nema sve što biološki mozak ima- određene kemijske procese koji se pojavljuju samo u biološkim bićima. Dakle, umjetni stroj bi morao imati kemijski sastav neurona i živčanog sustava poput onog u biološkim organizmima da bi mogao i funkcionirati i djelovati kao mozak živog bića.

Ovo razumijevanje nas dovodi do zaključka da naposljetku takav umjetni 'mozak' koji je sastavljen od iste materije i funkcionira kao organski mozak, nema smisla zvati umjetnim. Ovime nam Searle nameće zaključak da ne može postojati misleći stroj, osim onog koji je načinila priroda te ga čovjek nikad neće moći načiniti umjetnim putem.

Kao odgovor na ovo treba reći da ipak još nije empirijski potvrđeno da svijest, misao i intencionalnost mogu proizaći *samo* iz organskih mozgova kao niti iz nekakvih drugih kemijskih spojeva i umjetnih materijala. Searle stoga ovo pitanje, kao što je slučaj i u ovom radu, ostavlja znanosti: „To je empirijsko pitanje, poput onog pitanja je li moguće dobiti fotosintezu iz nečega s drukčijim kemijskim svojstvima od klorofila.“[37] Zaista, i sam Searle govori da će znanost, ako ikad uspije organizirati eksperiment koji dokazuje postojanje svijesti, pokazati koliko su njegove tvrdnje bile pogrešne ili valjane.

No, cilj cijelog ovog misaonog eksperimenta jestda se odgovori na pitanje je li moguće reći da računalo ima intencionalnost i razumijevanje ako ima odgovarajuću vrstu programa. Searle odgovara da tonije moguće, jer sama manipulacija računala formalnim simbolima nema nikakvu intencionalnost, iako se može činiti da ju računalo ima- simulacija razumijevanja nije isto što i 'istinsko' razumijevanje.

Prema ovome, možemo zaključiti da se Bryson i Searle slažu u gledanju na umjetno inteligentni sustav kao na kompleksni alat, odnosno, da nema potrebe promatrati umjetnu inteligenciju kao moralnog pacijenta. No vratimo se na teze Zuckerberg koja daje zanimljiv još jedan odgovor i zaključak na cijelu problematiku kineske sobe.

---

[36]Arielle L. Zuckerberg, *Moral Agency and Advancements in Artificial Intelligence*, Claremont McKenna College, 2010., str. 22.
Pristupljeno 4.9.2019: https://scholarship.claremont.edu/cmc_theses/36/
[37]J.R.Searle, *Minds, brains and programs,*str. 422.

> „Ali važno je primijetiti da Searleov argument ne utječe na raspravu o moralnom statusu umjetne inteligencije. Kada je racionalni agent konfrontiran sa strojem koji je sposoban proći Turingov test, njezin moralni stav prema stroju mora biti baziran na njezinoj interakciji s njim, budući da određujemo naš moralni stav prema drugim bićima na ovaj način. Nije bit u tome može li agent osjećati bol, moramo moralno djelovati kao da osjeća, budući da ne zahtijevamo da agent dokaže da stvarno osjeća bol naspram naprostog ponašanja koje za namjeru ima 'simulaciju' iskustva boli.“ [38]

Ovdje vidimo zanimljivu dvojbu između Searle i Zuckerberg, s jedne strane imamo zahtjev za dokazivanjem svijesti u umjetnom sustavu, a s druge zahtjev da se ponašamo prema umjetnom sustavu kao da ima svijest. Razumljivo je da Searle želi dokaze o svijesti, no Zuckerberg ističe da se ne mogu producirati takvi dokazi, stoga je razumno ophoditi se prema umjetno inteligentnom sustavu kao da ima svijest jer ipak to radimo sa svim bićima na osnovi intuicije. Na prvu bismo se ruku mogli složiti sa Zuckerbergovom logikom, no moramo se čuvati antropomorfizacije umjetnih sustava uzevši u obzir ulogu umjetne inteligencije u gospodarstvu, trgovini i ekonomiji - stvar postaje malo kompleksnija.

Najrazboritije bi bilo pokušati utvrditi stvarno postojanje svijesti ili iskustva boli (ili nečega sličnog) u slučaju umjetno inteligentnijh sustava ako je to ikako moguće. Nema potrebe davati stroju status moralnog pacijenta ako on jednostavno ne trpi posljedice našeg djelovanja slično kako trpe druga bića. Davanje prava umjetnim sustavima znači ograničavanje čovjekove uporabe tih sustava što bi sa sobom nosilo vlastite probleme.

Na ovu tvrdnju bi se moglo prigovoriti da je forma specizma[39], no moramo imati na umu da umjetni sustavi jednostavno *nisu* poput bioloških bića. Logično je pretpostaviti da biološka bića imaju nekakva mentalna i emotivna stanja jer i čovjek kao biološki organizam ih ima i također ih može prepoznati u drugim bićima na ovaj ili onaj način. S umjetnim sustavima to nije nužno slučaj, te nema potrebe upuštati se u neku formu animizma.

Jasno je da, ako se pokaže da je dokazivanje svijesti nemogući zadatak, a postoje dobri neizravni dokazi koji upućuju da biumjetni sustav mogao imati svijest,onda bi *morali* djelovati s pretpostavkom kao da ju *zbilja* ima te tretirati umjetnu inteligenciju kao moralnog pacijenta s odgovarajućim pravima i potrebama.

Da zaključimo, iznesen je Searlov argument koji tvrdi da umjetna inteligencija ne može imati nešto što bi smo nazvali sviješću, unatoč tome što bi umjetni sustav mogao ostavljati dojam

[38] A.L. Zuckerberg, *Moral Agency and Advancements in Artificial Intelligence* str. 23.
[39]Specizam u smislu kojem ga upotrebljava Peter Singer.

da ima. No, kao i što sam Searle kaže, treba se provjeriti znanstvenim metodama može li u anorganskim sustavima postojati svijest kakvu nalazimo kod živih bića. Ovdje valja kritički primijetiti da se pod „znanstvenim metodama“ misli isključivo na metode prirodnih znanosti. No, svijest se ne da znanstveno objasniti, „niti u jeziku fizike, niti u jeziku kompjutora“.

3.3. Teorije svijesti

Ovim se dijelom nastoji dati replika na tvrdnju da bi umjetna inteligencija samo simulirala svjesnost i *donekle* pokazati kako s filozofskih, tako i sa prirodoznanstvenih stajališta, da bi umjetni sustav mogao „biti svjestan“. Budući da postoji puno različitih teorija o svijesti, u ovom dijelu rada objasnit će se one teorije prema kojima bi i umjetni sustav, tj. umjetni opće inteligeni ili umjetni super-inteligentni sustavi, mogli razviti svijest. Ovime nastojimo pokazati da je moguće i da umjetni sustavi mogu imati nekakvu formu svijesti koja nije simulirana. Ove prirodoznanstvene teorije imaju svoje prednosti i nedostatke u objašnjavanju fenomena svijesti koji se neće dublje analizirati, osim kada je to bitno za razumijevanje tih teorija.

3.3.1.Emergencija

„Emergencija“ je jedan od središnjih pojmova konekcionizma čija se glavna ideja sastoji u prijenosu neurobiološki istraženih moždanih funkcija na informacijske sustave. Stoga se i govori o „bioinformatici“ ili „neuroinformatici“. Ova teorija uma govori da određeni entiteti posjeduju svojstvo emergencije. To svojstvo, ugrubo rečeno, znači da nekakva nova supstancija ili svojstvo može „proizaći“ iz osnovnih entitetaili iz njihove interakcije. Te nove supstancije i svojstva drugačija su u odnosu na one iz kojih su proizašla, stoga ih se ne može reducirati na njih, ona nisu nekakav dodatak i ne mogu zamijeniti ili promijeniti funkcije i interakcije osnovnih entiteta. U našem slučaju svijest bi bila emergentno svojstvo mozga i ona bi proizlazila iz neuroloških procesa, ali ona ne bi bila svediva na neurološke procese. Pojam emergencije često je ispunjen različitim sadržajima. Kritički treba primijetiti da nešto kvalitativno novo ne može nastati putem povećanja kvantitativne kompleksnosti.

Epistemološke pretpostavke emergentizma su donekle prihvaćene dok one ontološke, koje se odnose na svijest nisu, ali otvorene su za raspravu. Neki filozofi tvrde da postoje dovoljni razlozi za pretpostavljanje da svijest, intencionalnost, itd. su ontološki emergentna svojstva jer se intrinzična kvalitativna i intencionalna svojstva našeg iskustva čine fundamentalno različite

od naravi fizičkih i bioloških svojstava i procesa. Uz to, tvrde da naše iskustvo vlastitih namjera i odluka sugerira oblik nekakve 'direktne' makroskopske kontrole nad našim generalnim ponašanjem. Ova vrsta opće kontrole se, dakle, ne može reducirati na sumu individualnih procesa u relevantnim dijelova mozga.[40]

Ovom teorijom možemo reći da je fenomen svijesti nešto što prirodno proizlazi iz funkcija i sastava mozga. Pojavu *emergencije* vidimo posvuda u prirodi, i u ljudskim društvima, i kreacijama te je donekle i logično pretpostaviti da je i svijest nastala kroz emergenciju ili na sličan način. Smatra se daje zbog tog pogleda na ovaj fenomen kao na nešto jedinstveno i mistično teško adekvatno i objektivno raspravljati o svijesti umjetne inteligencije. Ona u ovom slučaju postaje još jedna sasvim prirodna pojava koja, sasvim realno, može nastati ili replicirati i u umjetnim sustavima u nekakvom drugačijem obliku. Ovdje ponovo možemo uputiti kritičku objekciju kako bi se najprije trebala moći shvatiti cijelu kompleksnost ljudskog mozga da bi bila moguća rekonstrukcija unutar umjetne kopije. Ali o tome nemamo znanja. Ako se polazi od toga da se sve ljudske, posebice kognitivne sposobnosti mogu objasniti složenim, paralelnim i fleksibilnim umrežavanjem jednostavnih elemenata i principjelno tehnički imitirati, onda na djelu imamo redukcionizam i razne oblike „izama“, kao što su fizikalizam, biologizam i sl. to ćemo pokazati na primjeru sljedeće teorije svijesti.

3.3.2.Kvantna teorija svijesti

Ova teorija svijesti je prilično opsežno i kompleksna prirodnoznanstvena teorija, stoga će se iznijeti samo osnovne teze teorije.

Cijela se teorija sastoji od grupe hipoteza koje tvrde da se klasičnom mehanikom ne može objasniti svijest te da bi ju fenomen kvantnih mehanizama poput kvantnog sprezanja i superpozicija mogli objasniti jer imaju važnu ulogu u moždanim funkcijama. Postoje razne teorije kako bi kvantni mehanizmi mogli biti upleteni u nastanak svijesti, ali te tvrdnje nisu dokazane. Jedna od mnogih hipoteza je Penroseova u kojoj je predložio da objektivna redukcija, tj. određeni oblik kolapsa valne funkcije[41], predstavlja ne-komputacijski utjecaj na geometriju vremena i prostora iz koje može proizaći svijest. Postoje također i razni prijedlozi

[40] Timothy O'Connor and Wong, Hong Yu Wong, „Emergent Properties“, Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy,* Summer 2015 Edition.
Pristupljeno: 7. 4. 2019:https://plato.stanford.edu/entries/properties-emergent/

[41]Sažimanje valne funkcije koja ima oblik superpozicije više svojstvenih stanja u jedno stanje koje se opaža u eksperimentu. (Institut za hrvatski jezik i jezikoslovlje)

kako načiniti eksperimente kojima bi se ovo moglo dokazati ili opovrgnuti, ali za sada, ne postoje načini da se ti eksperimenti provedu.

Dakle, ove hipoteze u suštini tvrde da su svijest i materijalna realnost komplementarni aspekti iste stvarnosti. Jasno je da se kvantni procesi događaju u realnosti, pa tako i u biološkim sustavima i u mozgu. Moguće je da događaji na kvantnoj razini uzrokuju promjene u molekulama mozga, time mijenjajući stanje neurona što dovodi do većih promjena. No, pitanje je imaju li ti procesi ikakav učinak i jesu li relevantni za one aspekte moždanih radnji koje imaju veze s mentalnim aktivnostima.

Razlog zašto su se neki autori, predstavnici fizikalizma, okrenuli kvantnoj teoriji za objašnjenje svijesti i svjesnih slobodnih izbora jest nasumično svojstvo kvanta čime se mogu objasniti slobodna odlučivanja u determiniranom svijetu.[42] Naime, fizikalizam polazi od teze da postoji sličnost između kvantnofizikalnih fenomena i fenomena svijesti.

Ova teorija donosi zanimljive teze jer bismo, prema njoj, uz pomoć kvantne fizike mogli saznati kako nastaje i kako svijest funkcionira. Budući da govorimo o kvantnoj razini, znači da fenomen svijesti ne bi nužno bio ograničen samo na biološke organizme, već bi se mogao pojaviti i u kompleksnim umjetnim sustavima. Ovaj pristup suočava se sa teorijskim i praktičkim problemima. Kvantnofizikalni procesi ne mogu se predvidjeti na najelementarnijoj razini, dakle ne mogu se opisati algoritimički. Tehnološke poteškoće očituju se kao poteškoća izoliranja kvantnog sustava u odnosu na okolni svijet. Također je problem što je ovaj pristup redukcionistički, jer se mentalni procesi poistovjećuju s funkcijama.

3.3.3.Intergrirana teorija informacija

Ovu je teoriju svijesti (*Intergrated Information Theory*) predložio neuroznanstvenik Giulio Tononi i ona, za razliku od drugih teorija, uzima svijest kao činjenicu te pokušava matematički objasniti što ono jest i zašto je povezano s određenim fizikalnim sustavima. Ova teorija isto tako promatra svijest kao nešto što se prirodno nalazi u mnogim živim bićima, tj. kompleksnim sustavima. Naravno, ovakvi eksperimenti i istraživanja svijesti su tek na početku te se sama metodika ovog polja treba razvijati, no smatra se bitnom jer se u njoj vidi mogućnost potvrde ili opovrgavanja tvrdnje o svijesti umjetne inteligencije.

[42] Harald Atamanspacher , Quantum Theory and Consciousness: An Overwiev with Selected Examples, Discrete Dynamics, *Nature and Society* 1/ 2004., str. 57-73.Ovdje str. 52.
Pristupljeno 4.9.2019:http://dx.doi.org/10.1155/S102602260440106X

Prema ovoj se teoriji mogu dati mjerenja ili predvidjeti je li neki sustav svjestan, koja je njegova razina svijesti i kakvo iskustvo taj sustav ima. Ima li neki fizikalni sustav svijest, određuje se na osnovi njegovih kauzalnih svojstava. Prema ovoj teoriji svaki kompleksniji sustav ima broj (oznaku Φ) koji nam kazuje stupanj integracije sustava i daje informaciju o mjeri svijesti tog sustava. Svaki sustav koji ima integriranu informaciju veću od nule smatra se svjesnim. Bilo kakva integracija ima svoje vlastito iskustvo sebe, stoga se i smatra da je svijest intrinzično, fundamentalno svojstvo bilo kakvog fizikalnog sustava.[43]

Ova teorija definira osnovna svojstva iskustva, a time i svijesti. Drugim riječima, ona nastoji dati odgovor na upit kakva bi trebala biti bića da bi imala svijest. Teorija polazi od nekoliko aksioma. Prvi je da je svijest realan intrinzičan fenomen neovisan o vanjskim promatračima. Ovo vodi do toga da biće nužno mora postojati, da može djelovati na sebe, na realnost van sebe, ali i okolina mora moći djelovati na njega. Drugi aksiom odnosi se na „kompoziciju" a to znači da je svijest strukturirana i sastoji se od raznih fenomenoloških distinkcija. Primjerice, unutar jednog iskustva će se razlikovati knjiga, bijela boja, bijela knjiga, lijeva strana, bijelu knjigu na lijevoj strani, itd. Iz ovoga slijedi da i biće mora biti strukturirano, da se mora sastojati od nekih osnovnih elemenata koji mogu imati djelovanje jedno na drugo ili na cijelo biće. Strukturiranost omogućava da osnovni elementi formiraju mehanizme višeg poretka i da više mehanizama formira strukturu.

U svojoj recenziji rada *„From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0"* Tononi Giulionavodi nekoliko pretpostavki o fenomenu svijesti koje koristi kako bi uopće mogao provoditi svoja istraživanja.

Prva takva pretpostavka govori da svijest postoji kao fenomen i da je specifična. Svako iskustvo je posebno zbog specifičnih fenomenoloških distinkcija što ga čini različitim od drugih iskustava. Iz ovoga slijedi da je i biće jedinstveno. Nadalje, tvrdi se da je svijest jedinstvena te da se svako iskustvo ne može reducirati na pojedinačne, razdvojene fenomenološke distinkcije. Isto tako i biće ne smije moći biti reducirano na svoje komponente. Uz ovo navodi da je svijest konačna i da se svako iskustvo sastoji od određenog

[43]Masafumi Oizumi, Larissa Albantakis, Giulio Tononi., From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0, *PLOS Computational Biology*, 2014. str. 1,2. Pristupljeno 9.4.2019:https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003588

seta fenomenoloških distinkcija prema kojima se uviđa da iskustvo ne može imati manje sadržaja i ima određeno vrijeme trajanja.[44]

Čini mi se da bi nam ovakvo viđenje svijesti donekle moglo pomoći pri razmatranju mogućnosti postojanja svjesnog umjetnog sustava, kao i u razumjevanju načina funkcioniranja jedne strojne svijesti.

3.3.4.Neuralne osnove svijesti

Ovdje će se referirati naznanstvena istraživanja kojima se bavi neuroznanstvenik Christof Koch te će se razmotriti kako njegovi zaključci utječu na promišljanje problema svjesnosti umjetnih sustava.

Naime, Koch se zalaže za tezu da je svijest fundamentalno svojstvo umreženih entitetai da ne može biti izvedena iz bilo čega drugog, zato što je jednostavna supstanca: „Svijest dolazi s organiziranim komadima materije. Imanentna je u organizaciji sustava. Ona je svojstvo kompleksnih entiteta i ne može se reducirati na radnje elementarnih svojstava.“[45] Dakle, svijest proizlazi iz kompleksnih umreženih sustava (emergentizam) te je *fundamentalno* svojstvo živih tvari.

Zbog toga ljudi i životinje, pa i kukci, imaju svijest, posebno iskustvo kako je biti upravo to što jest u različitim stupnjevima i načinima.[46] Koch je svoju tvrdnju jednostavno objasnio:„Ti i ja se nalazimo u kozmosu u kojem bilo koji i svi sustavi ususstavu dijelova koji su u međusobnoj interakciji posjeduju u nekoj mjeri svijesti. Što je veći i umreženiji sustav, veći je i stupanj svijesti. Ljudska svijest je puno istančanija nego pasja svijest jer ljudski mozak ima dvadeset puta više neurona nego mozak psa i više je umrežen.“ [47]

Koch ovdje govori o živim bićima jer su ona predmet njegovih istraživanja, no ako pogledamo karakteristike koje su nužne za svijest, prema Kochu, ondase može zaključiti da bi nekakva vrsta svijesti mogla postojati i u umjetno inteligentnom sustavu. Ovo i sam Koch priznaje kao mogućnost te objašnjava i zašto je totako.

---

[44]Oizumi M,, Albantakis L., Tononi G.,*From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0,* str. 2,3

[45]Koch Christof, , *Consciousness Confessions of a Romantic Reductionist*, The MIT Press, Cambridge 2012., str 59.

[46]Ovdje se Koch referira na integriranu toeriju informacija GiuliaTononija.

[47] Ch. Koch, *Consciousness Confessions of a Romantic Reductionist,* str. 59.

> „Ako se svaki akson, sinapsu i živčanu stanicu moga mozga zamjeni sa žicama, tranzistorima i strujnim krugovima koji izvršavaju sasvim istu funkciju, moj um bi ostao isti. Elektronička verzija moga mozga možda će biti nezgrapnija i veća, ali s time da svaka neuronska komponentna ima vjernu silikonsku imitaciju, svijest bi opstala. Za um nije važna priroda tvari od kojih je mozak načinjen, bitna je organizacija tih stvari– način na koji su dijelovi sustava spojeni, njihove kauzalne interakcije. Ljepši način da se ovo kaže je, 'Svijest je neovisna od [bioloških] supstrata'.“[48]

Time Koch objašnjava zaštonije nužna točna kopija mozga s njegovim kemijskim procesima, kako je predložio Searle, nego je bitno replicirati stanje naboja ili bez naboja kakvo vidimo kod neurona i sinapsi, što nas dovodi do mogućnosti svjesnog umjetno inteligentnog sustava.

Dakle, svijest se *može* pojaviti u računalnim sustavima pod uvjetom da imaju dovoljno kompleksnu mrežu entiteta s nabojima. Naravno, sadašnji umjetno intelgientni sustavi nisu toliko kompleksni i pitanje je hoće li biti potrebe napraviti takav sustav koji će morati imati ovakvu vrstu kompleksnosti ako želimo da posjeduje opću inteligenciju.

Iz svega dosad navedenog u ovim poglavljima možemo naposljetku zaključiti da nije sasvim jasno kako svijest nastaje i što ona jest, no na nju ne možemo više gledati kao na nešto mistično i svojstveno samo živim bićima. Budući da ne možemo sa sigurnošću definirati svijest, ne možemo niti sa sigurnošću tvrditi da u kompleksnom sustavu kao što bi trebala biti umjetna opća inteligencija i super-inteligencija (možda čak i neke vrste uske umjetne inteligencije) ona *ne može* nastati. Ono što možemo napraviti jest donekle opisati neke karakteristike svijesti, ali iz njih ne proizlazi zaključak da se svijest može pronaći *isključivo* u živim bićima. Dok se ne dokaže suprotno, ostaje realna mogućnost, prema sadašnjim istraživanjima znanstvenika što sto su Koch i Tononi u umjetnim sustavima *može* nastati nekakav oblik svijesti što će se uzeti kao potvrdu da naša rasprava o moralnom statusu umjetne inteligencije ostaje smislena i vrijedna razmatranja jer smo pokazali da fenomen svijesti može postojati u umjetnom sustavu što povlači sa sobom i ostale etičke implikacije.

Nakon ovog zaključka o svijesti umjetno inteligentnih sustava, trebamo se zapitati kako bi svijest jednog računala mogla funkcionirati?

[48]Ibid., str. 59.

3.4. Problem antropomorfizacije umjetne inteligencije

Jedna od stvari na koje stručnjaci upozoravaju jest upravo problem antropomorfizacije umjetne inteligencije. Naime, kada se problematizira o umjetnim opće ili umjetnim super-inteligentnim sustavima i mogućnosti da razviju svijest, najčešće se misli da će svijest, način razmišljanja i ponašanje ovakvih sustava biti slična ljudskoj.

Antropomorfizacija ovih sustava može biti očekivana, zato što smo naviknuli visoku inteligenciju poistovjećivati s čovjekom jer je on ipak najinteligentnije biće koje smo ikad susreli, stoga nam je intuitivno misliti da će i drugi inteligentni sustavi funkcionirati slično nama. No, moramo razumjeti da visoka inteligencija ne vodi nužno do čovjekove vrste svijesti ili razmišljanja, stoga je važno ograditi se od takvog načina razumijevanja svijesti iz sustava.

Na ovu su sliku čovječnog umjetno inteligentnog sustava utjecali i mediji, a posebno žanr znanstvene fantastike. Postoji velika mogućnost da umjetno inteligentni sustav neće biti poput nas. Ako umjetna opća inteligencija ili umjetna super-inteligencija razvije svijest, ona će nam biti najvjerojatnije toliko strana da ju nećemo moći naprosto pojmitizato jer je ona u suštini– strojna.

Kako bi ove tvrdnje bile jasnije, predstavit će se usporedba koju je predložio Tim Urban, a čini se dobro opisuje problem i omogućava nam da si malo bolje približimo intuitivno razumijevanje funkcioniranje strane svijesti umjetnih sustava. Urban u svom eseju o UI revoluciji daje sljedeću usporedbu.

> „Ako mi date hrčka u ruku i kažete mi da me sigurno neće ugristi, to bi me vjerojatno veselilo. Bilo bi drago. Ako mi onda date tarantulu u ruku i kažete mi da me sigurno neće ugristi, vjerojatno bih vrisnuo i ispustio ju iz ruku i pobjegao iz prostorije i više vam nikada ne bih vjerovao. No u čemu je razlika? Ni jedan nije bio opasan ni na koji način. Vjerujem da odgovor leži u stupnju životinjine sličnosti meni.“[49]

Naime, Urban govori da postoji nekakva biološka poveznica između hrčka i čovjeka, stoga osjećamo određenu empatiju prema nečem sličnom nama- drugom sisavcu. Čovjek može donekle razumjeti i poistovjetiti se s hrčkovim emotivnim stanjima i ponašanjem upravo zato jer funkcioniramo na sličan način kao i drugi sisavci. Razumijemo kako bi se hrčak, pas ili majmun mogao osjećati kada mu se pruži hrana i pažnja ili ako mu se nanese nepravda i zlo. Možemo pretpostaviti kako oni to percipiraju jer vidimo da odgovaraju na takve tretmane

[49]Urban Tim, *The AI Revolution: The Road to Superintelligence*, 1.22, 2015.
Pristupljeno 4.9.2019: https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html

slično kao i mi te time možemo i pretpostaviti kakva je svijest sisavaca. No, s paukom je drugačija priča. S njim nemamo bliže biološke poveznice te nam je njegova svijest potpuno strana jer je on *insekt*. Upravo ta stranost funkcioniranja i percepcija insektovog mozga i njegova ponašanjaje je ono što bi nas natjeralo da ispustimo pauka iz ruke, njegova svijest nam je nepojmljiva.

Umjetni opće i super-inteligentni sustavi bi bili, u našem slučaju, poput jednog pauka s ljudskom inteligencijom ili s inteligencijom većom od ljudske. Postavlja se pitanje, hoće li jedan takav sustav osjećati emocije –empatiju ili mržnju kao mi? Vjerojatno neće jer, kako je već napomenuto, ne postoji razlog zašto bi viša inteligencija vodila do razvitka ovakvih karakteristika. Ovi sustavi bi mogli imati osjećaje jedino ako ih mi implementiramo.

Dakle, teško nam je razumjeti kako funkcionira jedna strana svijest poput paukove, no teže će nam biti razumijevanje svijesti jedne umjetne opće ili super-inteligentnog sustava koji čak nije ni biološki organizam, nego *stroj*. Urban nas ovom usporedbom upozorava da bez obzira na izvanjsko ponašanje umjetnih opće i super-inteligentnih sustava. Koliko god ono nalikovalo na ljudsko, ono će u svojoj suštini imati nama jednu veoma *stranu* svijest te zato moramo izbjegavati antromprfizaciju umjetno inteligentnih sustava ipristupati veoma oprezno pri konstrukciji ovakvih računalnih sustava.

Iz svega je navedenog jasno da umjetna opća ili umjetna super-inteligencija najvjerojatnije neće misliti i osjećati kao čovjek te će njihov način interakcije sa svijetom, kao i sam pogled na svijet, biti bitno drugačiji od svega što smo do sad susreli. Zbog toga se smatra i da etički principi kojima se vodimo pri odnošenju s drugim živim bićima neće biti sasvim prigodni kada je u pitanju etički odnos prema umjetnim bićima.

Bez obzira kako ta svijest funkcionirala, hoće li umjetno inteligeni sustav osjećati ili ne, složili smo se s tim da ako umjetno inteligentni sustav posjeduje svijest, on postaje moralni pacijent. Isto kao što se slažemo da su hrčak i pauk moralni pacijenti bez obzira što je njihova svijest drugačija od naše. Jasno je da drugačije doživljavanje realnosti ne znači da biće ne zaslužuje da se prema njemu odnosi neetički.

Kako je primijetio Josef Weisenbaum malo znamo i o ljudskoj inteligenciji, te upozorava da smo pred novim etičkim pitanjima i mora se obratiti pozornost kako se sve to reflektira na poimanje o našem mišljenju i djelovanju. Nadalje, Eidenmueller ističe kako relevantni problemi nisu ni tehnički, ni matematički, nego etičke naravi. Etička refleksija o

umjetnoj inteligenciji pretpostavlja osvjetljavanje čitave sveze između tehničke civilizacije i etičkih mjerila. Čovjek ne stoji pred pitanjem značenja i odgovornosti pred njemu 'izvanjskom tehnikom', nego je također i u pitanju on sam.

4. ETIČKA REFLEKSIJA O UMJETNOJ INTELIGENCIJI

U ovom poglavlju ćemo pokušati analizirati koliko se principi i shvaćanja deontologije Tom Regana i utilitarizma Petera Singera mogu primijeniti na našu raspravu o moralnom statusu umjetne inteligencije. Također ćemo analizirati koji su mogući problemi pri primjeni ovih teorija u ovakvoj raspravi. Odabrala sam ove dvije etičke teorije jer mislim da nam one najbolje mogu pomoći u razmatranju moralnog statusa umjetno inteligentnog sustava. Započet ćemo analizu sa kratkim referiranjem na zakone i prava jer često serasprave u vezi etičkog razmatranja upravo svode na raspravu o mogućim pravima umjetno inteligentnih sustava jer se moralni status može predočiti uz pomoć zakona i kako je ona donekle slična raspravi o moralnom statusu životinja.

Naime, znamo da prema deontološkoj etici imamo dužnost etički se odnositi prema bićima s umom te ih uvijek trebamo promatrati kao svrhu. Možemo reći da se ova etika očituje kroz davanje zakonskih prava bićima jer to znači da smo prepoznali vrijednost tog bića kao i njegove potrebe te nas upozorava na moralne dužnosti koje imamo prema drugome. Davanje prava biću znači da je zabranjeno ili da se neće tolerirati ikakva diskriminacija i nanošenje zla tom biću. Također nosi ideju da se ne može tolerirati bilo kakvo *opravdanje* kršenja nečijih prava iz bilo kojeg razloga pa čak i za postizanje dobroga za razliku od utilitarizma. Davanje prava umjetno inteligentnim sustavima se može dovesti u usporedbu s davanjem prava životinjama. Unatoč tome što životinje nemaju sposobnosti koje ima umjetna inteligencija, njima svejedno omogućavamo zakonsku zaštitu. Stoga bi bilo zanimljivo vidjeti u kojoj mjeri možemo usporediti životinje i umjetno inteligentne sustave kada je riječ o pravima i dali se argumenti koji se inače koriste u prilog zaštite životinja mogu primijeniti na slučaj umjetne inteligencije.

Nadalje ću ukratko izložiti neke ideje koje zastupaju Tom Regan i Peter Singer kada se zalažu za prava i zaštitu životinja te uz pomoć njih pokušati opravdati ideju da i umjetna inteligencija zaslužuje imati status moralnog pacijenta.

Gledajući iz deontološke perspektive Tom Regan je odlučio argumentirati zašto životinje zaslužuju svoj moralni statustako što je pronašao osnovne zajedničke karakteristike čovjeka i životinjate izjednačio njihovu vrijednost. Kao i što čovjek ima potrebe, osjećaje, sjećanja i stalo mu je do vlastitog dobra, tako imaju i životinje. Naravno, te potrebe, osjećaji ili sjećanja se razlikuju, ali stvar je u tome da oni tvore jedinstveno *iskustvo života:*

„Sličnosti, između ljudskih bića koja najjasnije, najnekontroverznije imaju takvu vrijednost (na primjer, ljudi koji ovo čitaju), a ne naše razlike, su najvažnije. Ta stvarno krucijalna, osnovna sličnost je jednostavno ova: mi smo, svatko od nas, iskustvujući subjekt života, svjesno stvorenje s vlastitim interesima koji su nam važni kakve god da smo korisnosti drugima.“[50]

Nijedno iskustvo života ne može biti vrjednije, već se mogu *samo* razlikovati. Jedna verzija stvarnosti nije vrjednija od druge. Upravo to *iskustvo života* Regan smatra moralno vrijednim. Stoga Regan želi da se fokusiramo na tu zajedničku karakteristiku-*iskustvo života*, a ne na razliku u tim iskustvima. Usto također imamo i činjenicu da je životinjama stalo do vlastitog dobra, dakle imaju nekakve *interese*. Naime, u životinji se treba vidjeti subjekt koji posjeduje određeno iskustvo života i koji želi živjeti i ne želi patiti te na osnovu toga tvrdi:

„Osnovna vrijednost, prema tome, jednako pripada onima koji su iskustvujući subjekti života. Bez obzira pripada li to drugima–kamenju i rijekama, drveću i glečerima, na primjer– ne znamo i možda nikada nećemo znati.“[51]

Naposljetku, bića su *subjekti života* te time zavrjeđuju da ih se tretira kao svrhu ukoliko imaju osjećaje, sjećanja, potrebe i želju za vlastitim dobrom-iskustvo života. Iz navedenih informacija i citata vidimo da samo donekle možemo i umjetnu inteligenciju smatrati *subjektom života*. Regan spominje samo nekoliko stvari koje smatra da sačinjavaju iskustvo života (osjećaji, sjećanja itd.) i njih se do neke mjere može primijeniti na umjetnu inteligenciju, no ne iu potpunosti.

Ono što je problematično u ovoj raspravi o Reganovoj argumentaciji jest to što je donekle neodređena. Stoga, ako uzmemo Reganovu definiciju *subjekta života* doslovno onda možemo reći da umjetna inteligencija u kojoj nema potreba ili želje za vlastitim dobrom, na način kao što živa bića imaju, nije *subjekt života*i time nema potrebe moralno se odnositi prema njoj. No, ako uzmemo u obzir njegov cjelokupni stav, shvatit ćemo da te karakteristike nisu ono što je najvažnije u cijeloj njegovoj argumentaciji jer kao što vidimo Regan je sažeo svoj stav na: „…osnovna sličnost je jednostavno ova: mi smo svatko od nas iskustvujući subjekt života, svjesno stvorenje s vlastitim interesima…“[52] Drži se da se ovime nastoji reći da svako biće

[50] Tom Regan, A case for animal rights,*Advances in animal welfare science*, Washington, DC: The Humane Society of the United States.1986/87.,str.179-189. Ovdje str. 186.
Pristupljeno 4.9.2019:http://www.animal-rights-library.com/texts-m/regan03.pdf
[51] Ibid., str 186.
[52]Ibid., str. 186.

ima svoje iskustvo realnosti i to iskustvo je donekle zaokruženo nekakvim karakteristikama toga bića (osjećajima, sjećanjima, interesima i sl.), jest vrijedno samo po sebi.

Zato, ako bismo htjeli reći da jeumjetni sustav vrijedan kao biće i time moralni pacijent, trebao bi imati donekle svjesno ili zaokruženo iskustvo svog postojanja.

Ovakva deontološka argumentacija-traženje osnovnih zajedničkih karakteristika se donekle podudara s utilitarističkim razmatranjem etičkog i ontološkog statusa životinja prema Peteru Singeru.

Utilitaristička teorija općenito zahtijeva od nas da, pri donošenju etičkih odluka, razmislimo na koje i koliko individua naša odluka utječe te koji izbor će dati najbolje rezultate. Moralna dužnost je onda u odabiru opcije koja će pružiti najbolji omjer između ugode i patnje bića na koja moja odluka utječe. Prema Singerovoj poziciji utilitarista, jednostavno rečeno, cilj je smanjiti količinu zla ili nelagode i povećati količinu sreće i ugode. Ostvarenje tog cilja podrazumijeva obzirnost prema interesima drugog bića bez obzira na vrstu kojoj pripada.

Nadalje, ono što je privlačno kod ove etičke teorije jest to što u sebi sadrži egalitarizam. Svačiji interesi su važni i jednako vrijedni i diskriminacija među njima nije dopuštena jer ustvari nema dovoljno dobre osnove za to. Utilitarizam nam kazuje da možemo etički razmatrati biološka i umjetna bića pod uvjetom da su svjesna, čime je povećan krug etičkog razmatranja. U slučaju da se ne slažemo s ovim tvrdnjama i pokušamo utemeljiti razloge etičkog statusa u nečemu drugom, Singer nas upozorava da time zapadamo u tzv. specizam, odnosno, postupamo rasistički prema drugim vrstama osim ljudske. Ovu distinkciju između vrsta je također i Regan dovodio u pitanje.

Međutim, problem ove teorije jest u tome što ne postoji ta intrinzična vrijednost bića, jedino što je bitno su interesi tih bića i zadovoljenje tih interesa. Samo biće se gubi iz vida, pa samim time i njegova vrijednost. Ono što ostaje jest shvaćanje da biće ima svijest jer tada ima i interese vezane za svoju dobrobit. Ovdje, kao i kod Toma Regana, može se uvidjeti postojanje tvrdnje da nema svako biće istu vrstu svijesti, pa time i ne može se reći da svako biće pati na jednak način i u jednakoj mjeri.

Kada govorimo o umjetnoj općoj i super-inteligenciji razmatramo bića koja bi posjedovala veliki intelekt, ali i veoma drukčije interese od životinja i ljudi. No, ako ćemo se voditi deontološkim ili utilitarističkim pristupom u etičkom razmatranju umjetne inteligencije, onda ćemo se složiti da su oni jednako važni kao i interesi živih bića (pod uvjetom da umjetna

inteligencija ima svijest koja može imati iskustvo i interese). U ovom trenutku može se samo nagađati o tome što bi mogli biti interesi ili kakva bi bila iskustva umjetnih sustava, te kakva prava i odnosis ljudima bi bili moralni za ovakve sustave. Na problem o interesima također se nadovezuje i pitanje o tome hoće li ovakvi sustavi moći patiti više ili manje od drugih živih bića? Dolazi se do raznih pitanja kao što je kako definirati patnju, pa i prepoznati da umjetni sustav doista doživljava nešto što bi se moglo nazvati patnjom?

Ono što je bitno za primijetiti jest to da obje etike zahtijevaju konzistentnost u pogledu sadržavanja određene, osnovne karakteristike koju biće mora imati da bi imalo ikakav moralni status, a time i prava. U ovim slučajevima to je svijest ili biti *subjekt života*. Vidimo da se Singerovi i Reganovi argumenti mogu primijeniti na slučaj umjetno inteligentnog sustava te bi se moglo zahtjevati da se, ako umjetni sustav zadovolji zadane kriterije, smatra u krajnjem slučaju, moralnim pacijentom kao što se smatra i životinje. Prema ovome umjetno inteligentne sustave se nebi smjelo smatrati nešto kompleksnijim alatom tj. predmetom, nego bi im se morala priznati određena prava i zabraniti njegovo ugrožavanje ili zlostavljanje.

4.1. Ontološki i epistemološki problemi

Ove etičke teorije, iako korisne u razmatranju moralnog statusa umjetne inteligencije, u sebi sadrže nekoliko bitnih pitanja i problema na koje valja obratiti pozornost. Pitanja koja postavljamo u vezi moralnog statusa umjetne inteligencije nas dovode ido etičkih pitanja vezanih za samog čovjeka i percepciju čovjeka kao bića s moralnim značajem.

Postoji nekoliko problema s pristupima deontologije i utilitarizma prema promišljanju etičkih problema vezanih za umjetnu inteligenciju poput *problema primjene,* odnosno, što bi se podrazumijevalo pod poštivanjem prava umjetnih bića? Ali uz njega imamo i one kritičnije poput opravdanja etičkog razmatranja tj. zašto bi se uzele navedene karakteristike bića kao one bitne za određenje moralnog statusa? Također postoje i drugi problemi poput jednostavne činjenice da današnji umjetni sustavi nemaju svijest niti mogu patiti čime ova rasprava postaje suvišnom ili preuranjenom. No, htjela bih napomenuti da se unatoč tome treba baviti ovim pitanjem jer, kako sam već prije predložila u radu, postoji mogućnost da umjetni sustavi razviju svijest kakva god ona bila.

Postoji opcija da se snize kriteriji kao što su to učinili Luciano Floridi i J.W. Sanders[53] koji su predložili da se uzmu u obzir svojstva poput sposobnosti interkacije, prilagođavanja i autonomije pri donošenju odluka. No, čak i ovdje dolazi se do problema jer npr. dojenčad nemaju sasvim sva ova svojstva, no svejedno im dajemo važnost, a time i određeni moralni status. Ovdje se dolazi do idućeg problema *marginalnih slučajeva*. Naime, ako odredimo svojstva koja određuju moralni status i ako ih svi ljudi i druga bića nemaju u svakom trenutku, implicira li to da nisu vrijedna moralnog razmatranja? Za primjer uzmimo opet malenu djecu koja ne zadovoljavaju uvijek neke kriterije, znači li to da ćemo ih izostaviti iz etičkog razmatranja?

Također, postavlja se pitanje treba li se stavljati naglasak na inteligenciju umjetnih sustava kako bi im se mogao pridati moralni status. Tu se dolazi do važanog problema što razlikuje ljudsku od umjetne inteligencije. Filozof Julian Nida-Ruemelin navodi kako umjetna inteligencija ne posjeduje intenciju, a to je važno obilježje ljudske svijesti. Weisenbaum, informatičar i profesor na Massachusetts Institue of Technology (MIT) u svojoj knjizi „*Moć kompjutora i nemoć uma*“, dao je kritiku pojma inteligencije sadržanom u pojmu „umjetna inteligencija“ sa svrhom da ukaže na besmislenost govora o inteligenciji koja bi bila mjerljiva i to neovisno o svakom odnosnom sustavu kao što su obitelj, individuum, kultura. Drugim riječima, pojam inteligencije se ne može izolirati od drugih ljudskih sposobnosti.

Teško se i složiti oko toga koje svojstvo bića zapravo ima moralni značaj i kako dokazati da neko biće ima upravo to svojstvo. Trebali se složiti da je za naš moralni kriterij potrebna upravo svijest? Moj odgovor na ovo bi bio da je svijest odlika kojom možemo razlikovati svjesni stroj od običnog stroja i time se sačuvati od animizma i ograničavanja čovjeka na mjestima za koje nema stvarne potrebe. Dakako, ne moramo se nužno složiti s ovim, jer na osnovu toga oko kojeg se svojstva složimo da je relevantno za određivanje moralnog statusa, moguće je da ćemo intuitivno drukčije promišljati neki etički problem. Međutim,kada bismo se i složili oko relevantnog svojstva kao što je svijest (ili nekog drugog), čini mi seda ti kriteriji mogu biti veoma apstraktni i neodređeni poput Reganovog *subjekta života*.

Utilitaristička i deontološka teorija pokušavaju riješiti ove probleme na različite načine, no postavlja se pred njih problem održavanja konzistentnosti i moralne značajnosti vlastitih ontoloških kriterija. Vidimo da se pitanja koja inače pronalazimo u etici kao i u ove dvije

[53]Luciano Floridi, J. W. Sanders., On the morality of artificial agents, *Minds and Machines* 2004, Vol. 14, Issue 3, Kluwer Academic Publishers, 2004., str 349–379., Ovdje str. 7.
Pristupljeno 9.4.2019: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.722&rep=rep1&type=pdf

etičke teorije mogu povezati sa slučajem umjetnih sustava i odgovori na njihće imati utjecati na naš pogled na moralni status umjetne inteligencije.

4.2. Kako odrediti postojanje svijesti u umjetno inteligentnom sustavu?

Pretpostavimo da smo se ipak složili oko toga da bi umjetna inteligencija trebala imati svijest kako bi se mogla smatrati moralnim pacijentom. Kako ćemo pokazati da umjetni sustav uistinu ima takvo nešto? Ovdje se susrećemo s klasičnimf ilozofskim problemom drugih svijesti. Ovo je problematično dokazati sa sigurnošću kod životinja, pa i drugih ljudi, stoga kako će se dokazati svijest u računalnom sustavu?

Može se reći da se mora vjerovati vlastitoj intuiciji pri procjeni svjesnosti umjetno inteligentnog ustava, kako govori Zuckerberg, no neće svi biti zadovoljni ovakvim odgovorom kao što nije ni Searle, nego će se zahtijevati da ponudimo nekakav eksperiment koji bi ovakvo nešto mogao pokazati. Smatram da je s našim postojećim znanjem i definicijama svijesti nemoguće napraviti jedan takav traženi eksperiment. Dolazi se do problema kako ga adekvatno načiniti kada će se svakom predloženom eksperimentu uvijek moći pronaći mana upravo zato jer nije sigurno što ustvari svijest jest.

No, umjesto da se pokuša pokazati svijest u umjetnom sustavu, Robert Sparrow kazuje da bi se umjesto toga moglo pitati kada bi sustav mogao ispuniti ulogu čovjeka u moralnoj dilemi. Stoga nam on predlaže modifikaciju Turingova testa kojom bi mogli odrediti smatra li umjetno inteligentni sustav, u najmanju ruku, moralnim pacijentom.

Modifikacija koji Sparrow predlaže je slučaj medicinske trijaže u kojoj se donose odluke o životu i smrti pacijenata: “U scenariju kojeg predlažem, uslijed nestanka struje bolnički administrator je suočen s odlukom kojem od dva pacijenta na aparatima za održavanje života nastaviti davati struju. Može očuvati život samo jednog i njezina odluka neće utjecati na druge živote. Znat ćemo da su strojevi postigli moralni status usporedan s ljudskim kada zamijenimo jednog od pacijenata s umjetnom inteligencijom i dilema ostane ista. Odnosno, kada bismo mogli prosuditi da je razumno očuvati egzistenciju stroja prije nego život ljudskog bića. Ovo je 'Turingov trijažni test'.“[54]

[54]Robert Sparrow, The Turing Triage Test*, Ethics and Information Technology*, 6, Kluwer Academic Publishers Hingham, MA, USA (4/2004), str. 203–213. Ovdje str. 204.
Pristupljeno 9.4.2019:https://link.springer.com/article/10.1007%2Fs10676-004-6491-2

Ono što Sparrow ovime poručuje jest da onog trenutka kada se u umjetno inteligentnom sustavu prepoznaju karakteristike koje ima i čovjek poput racionalnost, autonomnost u donošenju odluka itd., ili u terminologiji Sparrowa, svijest, ambicije i projekti, onda će se morati razmotriti moralni status umjetne inteligencije. Stoga,Sparrow pokušava pokazati da odgovor na pitanje je li umjetna inteligencija moralni pacijent zapravo ovisi o tome je li riječ o moralnom djelovatelju.

Julian Nida -Rümelin odlučno odbacuje da se sve rečeno može etički zastupati jer strojevi ne mogu djelovati autonomno. Prema Nida - Rümelinu autonomija označuje ljudsku sposobnost voditi se u svom prosuđivanju i djelovanju umnim razlozima. Umjetno inteligentnom sustavu nedostaje razboritost ili moralna prosudbena moć, koja je nužna u donošenju odluka, posebice u dvojbenim situacijama.[55] Već je informatičar Weizbaum svojom kritikom „rane uporabe" umjetne intelgiencije u vijetnamskom ratu istaknuo: „Divovski kompjutorski sustavi u Pentagonu (...) nemaju autore". Kao proizvod mnogih programera koji ne poznaju cjelinu sustava, ti sustavi ne shvaćaju pitanja o ispravnom i pogrešnom. Istaknuo je kako „ne poznajemo mogućnost učiniti kompjutore razboritim i zato ne bi trebali na kompjutor prenijeti zadaće čije rješenje zahtijeva razboritost" .[56]

Naizgled je ono što se ustvari traži od UI tehnologije, kako su primijetili Luciano Floridi i J.W.Sanders, to da umjetna inteligencija pokaže svjesno djelovanje[agency] u smislu u kojem ga čovjek pokazuje kako bi ga se smatralo moralnim pacijentom. No, onda ga se ne može promatrati samo kao pacijenta, nego i kao moralnog djelovatelja. Ako umjetna inteligencija može donositi etičke i druge prosudbe poput čovjeka, u tom bi slučaju efektivno imao isti moralni status kao i čovjek te time imao vlastita prava kao što i ljudsko biće ima svoja prava.

U drugu ruku, postoji i verzija prema kojoj se može tvrditi da ako umjetna inteligencija može donositi etičke prosude te time jest moralni djelovatelj, ali i dalje nije i pacijent jer npr. ne osjeća bol ili jer je i dalje umjetan sustav. Tvrdeći da biće može biti moralni djelovatelj, ali ne i pacijent je izravno kršenje moralnih načela kojima se vodi kada je sam čovjek u pitanju, što je nepravedno i nemoralno. Ako neko biće pokaže svjesnost i samostalnost u rasuđivanju [agency] koju ima čovjek (kojeg se smatra moralnim djelateljem, što uključuje i to da je moralni pacijent), tada se mora priznati da to drugo biće zaslužuje isto takvo etičko

[55]Vidjeti šire: Nathali Weidenfeld, Julian Nida-Rümelin, *Digitaler Humanismus: Eine Ethik fuer das Zeitalter der kuenstlichen Intelligenz*, Piper Verlag, Muenchen, 2018.
[56]Joseph Weizbaum, *Die Macht der Computor und die Ohnmacht der Vernunft*, Frankfurt am Min, 1978., str. 300

razmatranje. U suprotnome, prema mom mišljenju, spadamo nanovo u Singerov specizam, jer čini mi se da ono što stoji iza takvih argumenata je, ustvari, čovjekova predrasuda prema umjetnom biću i čovjekova potreba da bude poseban među drugim bićima.

No, upravo su utjecaji tih predrasuda ono što bi otežalo izvršenje trijažnog testa, ako će se prema njemu pokušavati prosuditi o moralnom statusu umjetno inteligentnog sustava. Naime, čini mi se da ovaj trijažni test uključuje individue koje su ili veoma otvorenog uma ili se nalaze negdje u budućnosti kada su ljudi već donekle odbacili predrasude prema tehnologiji poput umjetne inteligencije. Stoga bi trebalo konstruirati verziju trijažnog testa koja bi se mogla primijeniti u današnje vrijeme, uzimajući u obzir navedene utjecaje.

4.3. Dokazivanje boli u umjetno inteligentnom sustavu

Nakon rasprave o moralnom statusu umjetnih sustava temeljenom na svijesti, trebalo bi se osvrnuti i na mogućnost da umjetni sustavi osjećaju nekakvu vrstu patnje. Ovdje se nailazi na pitanje kako je moguće da stroj ili računalni sustav iskusi patnju? Što uopće jest patnja i kako saznati da je umjetni sustav može iskusiti? Nailazi se na isti problem kao i s dokazivanjem svijesti. Moglo bi se nanovo raspravljati o ovim pitanjima, no umjesto toga ću ponuditi ulomak iz eseja Daniela Clement Dennetta koji kvalitetno upotpunjuje čitavu raspravu. Naime, pri kraju eseja "*Why You Can't Make a Computer That Feels Pain",* Dennett navodi da ovisno o našoj intuiciji o tome što jest bol zaključuje seosjeća li uistinu neko biće patnju ili ne:

> „Preporučam da ne pokušavamo očuvati ovu intuiciju, ali ako se ne slažete, koju god teoriju produciram, koliko god deskriptivna ili elegantna bila, neće biti zadovoljavajuća teorija boli, nego samo teorija onoga što sam samovoljno odlučio nazvati boli. Ali ako, kao što sam tvrdio, intuicije koje bismo morali poštivati ne formiraju dosljedan sustav, ne može postojati prava teorija boli i time nijedno računalo ili robot ne bi moglo utjeloviti pravu teoriju boli, što bi morali napraviti da bi osjetili realnu bol."[57]

Dakle, budući da postoji problem definicije boli, dolazi se do pitanja dali je onda moguće načintit umjetni sustav koji bi osjećao bol? Uz to, postaje i upitno hoće liovaj tip sustava zbilja iskusiti nekakvu vrstu neugode tekako to odrediti.

---

[57] Daniel Clement Dennett, Why You Can't Make a Computer That Feels Pain, *Synthesis*, Vol. 38, No. 3, Automaton-Theoretical Foundations of Psychology and Biology, Part I, Kluwer Academic Publishers, 1978., str. 415-456, Ovdje str. 449. Pristupljeno 4.9.2019. http://www.jstor.org/stable/20115302

Trenutno nema načina da se identificira iskustvo boli u umjetnom sustavu, stoga se tu procjenu može jedino temeljiti na ponašanju umjetne inteligencije no, Julian Nida-Ruemelin upozorava da se osjećaji ne mogu svesti na ponašanje. Naime, moramo razumijeti da *simulacija* bolnog iskustva, nije isto kao i iskustvo boli. Dirk Eidenmueller nam također kazuje da umjetna inteligencija nema kognicije, već se radi o pseudokognitivnom sustavom što znači da on „može samo simulirati ili imitirati svjesno ponašanje".[58] Stoga se ne može isključivo uzeti ponašanje umjetne inteligencije kao pokazatelja realnog iskustva boli, nego možda samo kao jedan, među mnogim, neizravnim dokazima bolnog iskustva. No čak je i ovo upitno jer je pitanje hoće li tehnologija moći proizvesti stvarno iskustvo, a ne naprosto simulaciju bolnog isukstva.

Naposljetku se opet dolazi do toga da je potrebna adekvatna definicija boli koja bi bila primjenjiva na umjetne sustave kako bi se mogli riješiti navedeni problemi. No, pod uvijetom da smo u mogućnosti načiniti umjetnu inteligenciju koja bi 'osjećala' bol može se primjetiti da bi, za razliku od svijesti koja bi mogla spontano nastati u sustavu, bolno iskustvo postojalo jedino ako čovjek pronađe način da ga namjerno implementira u sam sustav.

4.4. Etičnost inženjeringa umjetno inteligentnog sustava koji osjeća bol

No, sada se dolazi do idućeg pitanja, je li uopće potrebno i moralno izgraditi umjetni sustav sa sposobnošću da pati te kakve to implikacije podrazumijeva u svezi teme ovoga rada?

Bryson tvrdi da bi konstrukcija ovog tipa umjetnog sustava bila nemoralna : „Graditelji robota su etički obvezani–obvezani načiniti robote prema kojima vlasnici robota nemaju etičkih obveza." [59] Logično je složiti se s ovom tvrdnjom jer štiti umjetnu inteligenciju od iskustva patnje kao što i odrješava čovjeka od moralnih dužnosti prema umjetno inteligentnom sustavu.

Složila bih se da bi UI tehnologija trebala biti konstruirana na takav način jer onda nebi imali moralne obveze prema sustavu, niti bi sustav bio izložen neugodnim ili bolnim iskustvima. No, kao što sam već prije napomenula u radu, nismo sasvim sigurni u kojem će se smjeru tehnologija razvijati i postoji mogućnost da umjetni sustavi razviju svijest ili da se načine

[58]Dirk Eidenmueller, *Quanten – Evolution, Geist.Eine Abhandlung ueber Naturwissenschaft und Wirklichkeit*, Springer, Berlin-Heidelberg, 2017., str. 363
[59]J. Bryson, *Robots Should Be Slaves*, str. 10

sustavi koji mogu pericipirati bol, pa bi bilo mudro i preventivno voditi rasprave oko ovog problema.

No, zašto bi se željelo načiniti umjetno inteligentni sustav s mogućnošću percipiranja boli ili neugode? Ne bili to prouzročilo komplikacije, ne samo u čovjekovu odnosu s tehnologijom, nego i u ljudskom društvu općenito?

Pojedini znanstvenici i filozofi navode nekoliko razloga zašto bi željeli izgraditi umjeta nsustav koji bi imao implementiran osjet boli. Prvi je zato što se želi da naši roboti budu *empatični*. Bol je ključna u razvitku empatije prema drugome te bi uz pomoć nje mogli dobiti empatične strojeve. Budući da će razni umjetno inteligentni sustavi biti u dodiru s ljudima, važno nam je da oni mogu iskusiti bol kako bi istinski mogli suosjećati s ljudima i time bolje i sigurnije surađivati, i možda se čak i brižnije odnositi prema ljudima. Osobe kojima je potrebna pomoć ili njega ne bi htjeli simulacije empatije, nego bi htjeli znati da mogu vjerovati da ih stroj zbilja razumije.

Idućim argumentom se pokušava dodatno opravdati kreacija 'osjećajućih' umjetno inteligentnih sustava s logikom koju nekad primijenjujemo i na potomke. Dakle, možemo tvrditi da će iskustvo patnje na neki način obogatiti cjelokupna iskustva umjetnih sustava te bi se, kao i ljudskom, pronašla nekakva ravnoteža dobrog i lošeg iskustva. Dakle, nije *namjera* naškoditi sustavu, nego se na bol gleda kao na jedno od iskustava koje će umjetna inteligencija imati. Ovaj argument možemo upotrijebiti i u čovjekovom slučaju kada se debatira oko moralnosti stvaranja potomaka. Kada donosimo potomke u svijet, nama nije cilj *namjerno* djetetu izazvati patnju, patnja je nenamjerna, ali obogaćuje život tog djeteta. Stoga ako se složimo da je neetično konstruirati 'osjećajući' umjetno inteligentni sustav,onda moramo reći da je i za čovjeka imanje potomaka moralno neispravno.

Međutim, problem zapravo nastaje, kako Meghan Winsby kazuje, kada se shvati da patnja mora biti *namjerno* izazvana kod ovakvih sustava kako bi oni sami mogli učiti i kako bi se mogla testirati njihova ispravnost. Winsby nadalje tvrdi da vjerojatnost da će umjetni sustavi moći raditi besprijekorno bez prethodnog treninga i testiranja koje znači namjerno izazivanje patnje, je vrlo mala. Stoga jedino preostaje da se sustavu prezentira bolno iskustvo sve dok se ne postignu željene reakcije.[60]

[60] Meghan Winsby, *Suffering Subroutines: On the Humanity of Making a Computer that Feels Pain*, Western University, London, Association of Computation and Philosophy Annual Meeting: College Park, MD 2013., str. 4,5. Pristupljeno 4.9.2019: http://www.iacap.org/proceedings_IACAP13/paper_48.pdf

Uzmu li se u obzir riječi Meghan Winsby, složit ćemo se da bi takav tretman umjetnih sustava bio posve neetičan. Unatoč tome što bi empatični umjetni sustavi bili od velike koristi za čovjeka oni se, makar za sada, ne čine kao nešto esencijalno za čovjekov život te stoga namjerno izazivanje boli u bićima, bez prave potrebe, je nemoralno.

Vratimo se na početne uvjete pod kojima smo se složili da bi se umjetna inteligencija smatrala moralnim pacijentom- svijest i iskustvo boli. Naime, bol se ne može zadovoljavajuće definiratii nema načina da se sa sigurnošću identificirati iskustvo boli. Isto stoji i za fenomen svijesti. No, također se čini da bi iskustvo boli postojalo u umjetnom sustavu *jedino* ako ga čovjek uspije implementirati, za razliku od svijesti koja bi mogla spontano nastati u samom sustavu. Čini mi se da bi se iz ovog razloga moglo dati veću važnost postojanju svijesti, nego iskustvu boli umjetnog sustava ili čak odbaciti agrument da bi stroj morao osjećati kako bi ga se smatralo moralnim pacijentom. Naime, ako se stavi naglasak na osjećaj boli, onda se nameće pitanje koliko je ispravno reći da svjestan stroj bez iskustva boli nije moralni pacijent. Uz ovo, pokazalo se da bi konstrukcija takve umjetne inteligencije zapravo bila nemoralna jer bi se neugodno iskustvo moralo iznova izazivati u umjetnim sustavima. Iz ovih se razloga dovodi u pitanje validnost uvjeta da umjetna inteligencija osjeća bol kao pokazatelj moralnog statusa. Zato smatram da se naposljetku treba odbaciti drugi uvjet-bol kao nekakav indikator moralnog statusa umjetne inteligencije i zadržati samo onaj prvi-svijest.

Iz svega navedenog, može se zaključiti da se vjerojatno neće moći sa sigurnošću ustanoviti imaju li ovi umjetni sustavi svijest ili osjećaju patnju (osim ako se ne pojavi nekakvo otkriće u znanosti koje će nam to omogućiti). Svaki eksperiment ili argument koji sada možemo dati, a kojim pokušavamo dokazati ova svojstva u umjetnoj inteligenciji, neće biti savršen i moći ćemo pronaći podosta zamjerki, no smatram da je to u suštini nebitno. Nemamo, naposljetku, načina da takvo što detektiramo u drugom čovjeku ili životinjama te iste prigovore koje dajemo za mogućnost svjesnosti umjetne inteligencije možemo dati i za druga živa bića te se u suštini vraćamo na problem drugih svijesti. Stoga, u nemogućnosti saznavanja, jednostavno bismo trebali djelovati s pretpostavkom utemeljenom na dovoljnim (neizravnim) dokazima da dovoljno inteligentan i kompleksan sustav (umjetna opća i umjetna super-inteligencija) imaju formu svijesti.

Konačno, smatram da u slučaju ako procijenimo da bi ovi umjetno inteligentni sustavi zaista mogli imati svijest i osjećati patnju, tada bi trebali biti tretirati s poštovanjem jer iz dosad

iznesenih filozofskih analiza slijedi da bi se trebali pravedno odnositi prema svjesnoj umjetnoj inteligenciji.

ZAKLJUČAK

Kroz ovaj rad vidjeli smoda su umjetno inteligentni sustavi uistinu kompleksni i filozofski interesantni na mnogim razinama. Iz općih informacija o Umjetno inteligentnim sustavima i njihovim tipovima, te načinima kako će se možda u budućnosti postići opća umjetna inteligencija i super-inteligencija razumijemo da su oni značajni za samog čovjeka i da će njihove sposobnosti značajno utjecati na živote mnogih bića bilo na negativan ili pozitivan način.

Mora se imati na umu da ova umjetna bića neće percipirati svijet na isti način kao čovjek unatoč njihovoj inteligenciji i autonomnosti. Upravo zato što smo svjesni da ne smijemo antropomorfizirati umjetne sustave ne želimo bez razloga pripisivati moralni status nečemu što ga ne zaslužuje i time bezrazložno iskriviti našu percepciju realnosti i etičnosti. Joanna Bryson je jedna od osoba koja nas upozorava upravo na to. Ona je među mnogima koji smatraju umjetno inteligentne sustave kompleksnim alatima koje bi se tako i trebalo tretirati. Zastupnici ovog pogleda zahtijevaju neke pokazatelje ili dokaze svijesti, ili iskustva patnje u umjetnim sustavima kako bi ih se moglo smatrati bićima s moralnim statusom.

Kao odgovor na ovaj zahtjev moralo se istražiti jeli moguće da umjetni sustav posjeduje nekakav oblik svijest. Ukratko je ponuđeno nekoliko teorija svijesti; emergencija, kvantna teorija svijesti, intergrirana teorija informacija Giulija Tononija i zaključci istraživanja neuralnih osnova svijesti Christofa Kocha. One većinom promatraju fenomen svijesti kao na nešto što prirodno proizlazi ili nastaje u kompleksnim sustavima, no njihovu ispravnost još uvijek treba utvrditi. Međutim, iz nekoliko teorija svijestividi se da mogućnost svjesnosti umjetno inteligentnih sustava nije apsurdna te ju se treba uzeti u obzir pri etičkom promišljanju moralnog statusa umjetne inteligencije.

Nadalje, razmotrilo se i problematiku Searlove kineske sobe koja za cilj ima to da se zapitamo kako ćemo utvrditi ima li stroj zbilja svijest irazumijevanje na način koji ga ima čovjek. Searle zaključuje da je takvo što moguće jedino u biološkim sustavima te dokazivanje suprotnog prepušta znanosti. No, čak i njegov zanimljiv misaoni eksperiment nije zapravo postavio novo pitanje već ga je,prema mom mišljenju, samo preusmjerio sa živih bića na umjetne sustave. On nas, svejedno, upozorava na to da ne smijemo antropomorfizirati umjetnu inteligenciju te da bi trebali biti kritični kada pripisujemo svjesnosti umjetnim sustavima.

S druge strane, Robert Sparrow nam je ponudio drukčiji pogled na pitanje kako utvrditi postojanje svijesti svojom etičkom dilemom koju je nazvao Turingov trijažni test pokušavajući nam pokazati da se moramo donekle osloniti i na svoju intuiciju u slučaju UI tehnologije s obzirom na to da ne možemo dokazati svijest, a time niti osjećaj patnje. Nadalje, Daniel Dennett je predočio besmislenost zahtjeva da se u umjetnom sustavu pokaže osjećaj patnje ili boli jer se ne može konstruirati adekvatna teorija boli, pa time niti odrediti osjeća li nešto bol ili ne (što je također problematično i kad je u pitanju definiranje i dokazivanje svijesti).

Također smo vidjeli i da deontološka teorija kao i utilitarizam mogu usmjeriti kako promišljati moralni status umjetnog sustava. Isto tako, uz pomoć Toma Regana i Petera Singera, vidjeli smo da se čitavo pitanje donekle može dovesti i u analogiju s pravima životinja i njihovim moralnim statusom što nam također pomaže u usmjeravanju promišljanja. Vidjeli smo da i prema Reganu i Singeru donekle možemo primijeniti i deontološku i utilitarističku teoriju na etičko razmatranje moralnog statusa umjetne inteligencije pod uvjetom da se pokaže da su ti sustavi svjesni. No, i ove etičke teorije nisu savršene u pogledu svojih epistemoloških i ontoloških pitanja te mogu samo donekle pomoći u našoj dilemi. Mora se riješiti *problem primjene*, što bi značilo poštivati prava umjetno inteligentnih sustava i posvetiti se problemu *marginalnih slučajeva* u koje bi mogli biti uvršteni umjetno inteligentni sustavi. Ove teorije nisu sasvim adekvatne u svojoj primjeni na umjetne sustave što sugerira da ćemo trebati posve novu etičku teoriju kojom ćemo se moći adekvatno raspraviti ovaj problem.

Cijela problematika ovoga rada se dotiče pitanja ljudskog identiteta čija osjetljivost dovodi do toga da se ona nekad izbjegava ili zapostavlja pod izlikom da takva rasprava zapravo nema smisla te da je nastala pod utjecajem znanstveno fantastičnih filmova i knjiga. Smatram u konačnici da će čovjekova budućnost biti nadasve isprepletena tehnologijom zbog čega svaki aspekt tema vezanih za umjetno inteligentne sustave mora biti ozbiljno prodiskutiran jer ćemo u protivnom biti uvelike nespremni za monumentalne promjene koje donosi zora umjetne inteligencije. No, u ovom trenutku valjalo bi propitati kakve implikacije ima govor o umjetnog inteligenciji na razumijevanje čovjeka i uopće sliku čovjeka i njegova svijeta. Pritom je važno pokazati da pitanje da li se i kako čovjek razlikuje od umjetne inteligencije nije empirijsko nego filozofijsko pitanje, te da prirodoznanstveni i tehnici imanentni kriteriji i teorije pri pokušaju davanja odgovora zahvaćaju prekratko. Između ljudske inteligencije i

umjetne inteligencije postoji „ne samo graduelna, nego temeljna razlika“.[61] Naime, razmatrane prirodoznanstvene teorije imaju filozofijske premise – shvaćanje čovjeka analogno svijetu stvari, te u empirijskoj primjeni prvotno filozofijskih pojmova reduciranje njihova značenja, čime prekoračuju svoje granice i postaju „izmi“ (fizikalizam, biologiozam, funkcionalizam). Time, dakako, ni u kom slučaju ne dovodimo u pitanje značenje prirodoznansntvenih spoznaja, nego se kritički odnosimo prema naturalizmu, shvaćanju prema kojemu se „načelno svi fenomeni, uključujući mentalna i specijalno intencionalna stanja i procese, dakle i ljudsko djelovanje , mogu prirodoznanstvenim metodama u cijelosti opisati i objasniti“.[62]

[61]Werner Gitt, *Am Amfang war die Information*, Neuhausen, Stuttgart, 1994., str. 250.
[62] Nida- Rümelin Julian, *O ljudskoj slobodi*, Breza, Zagreb, 2007., str. 34.

LITERATURA

ATAMANSPRACHER, H., Quantum Theory and Consciousness: An Overwiev with Selected Examples, Discrete Dynamics, *Nature and Society* 1/ 2004., str. 57-73.

Pristupljeno 4.9.2019: http://dx.doi.org/10.1155/S102602260440106X

BOSTROM, N., *Superintelligence Paths, Dangers, Strategies*, Oxford University Press, 2014.

BRIGGLE, A., WAELBERS, K.,BREY, A. E. P., (ed.), *Current Issues in Computing and Philosophy*, Vol. 175. Department of Philosophy, University of Twente, The Netherlands, IOS Press, 2008.

BRYSON, J.J., Robots Should Be Slaves, *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and DesignIssues*, John Benjamins, 2010., str. *63-74*.

CALVERLEY, D. J., Android science and animal rights, does an analogy exist?, *Connection Science* 18:4, 2006., str. 403-417.

Pristupljeno 16.4.2019: https://www.researchgate.net/publication/252760244_Android_Science_and_the_Animal_Rights_Movement_Are_There_Analogies

CARR, J., *An Introduction to Genetic Algorithms Abstract*, Senior Project, 2014.

Pristupljeno 7.4. 2019:http://www.joinville.udesc.br/portal/professores/parpinelli/materiais/IntroductionGA.pdf

CHALMERS, D. J., *The Conscious Mind: In Search of a Fundamental Theory (Philosophy of Mind)*, Oxford University Press, Oxford NewYork, 1995.

DENNETT, C. D., Why You Can't Make a Computer That Feels Pain, *Synthesis*, Vol. 38, No. 3, Automaton-Theoretical Foundations of Psychology and Biology, Part I, Kluwer Academic Publishers, 1978., str. 415-456.

Pristupljeno 4.9.2019:http://www.jstor.org/stable/20115302

DREYFUS, H.L., *What Computors can't do?A Critique of Artificial Reason,* New York, 1972.

EIDENMUELLER, D., *Quanten – Evolution, Geist.Eine Abhandlung ueber Natur, Wissenschaft und Wirklichkeit,* Springer, Berlin-Heidelberg, 2017.

ERASSME, R., *Der Mensch und die „Kuenstliche Intelligenz". Eine Profilierung und kritische Bewertung der unterschiedlichen Grundauffassungen vom Standpunkt des gemässigten Realismus,* Philosophische Fakultät der Rheinisch-Westfälischen Technischen Hochschule, (disertacija), Aachen, 2002.

FLORIDI L., SANDERS, J. W., On the morality of artificial agents, *Minds and Machines*, Volume 14, Issue 3, Kluwer Academic Publishers, 2004., str. 349–379.

Pristupljeno 9.4.2019:http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.722&rep=rep1&type=pdf

FOGEL,D. B., *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence,* IEEE Press, Piscataway, 2006.

PATRICK, G. T. W., The Emergent Theory of Mind, *The Journal of Philosophy*, Vol. 19, No. 26, Journal of Philosophy Inc., Dec. 21, 1922., str. 701-708.

Pristupljeno23.3.2019: https://www.jstor.org/stable/pdf/2939801.pdf

GOERTZEL, B., PENNACHIN C., (ed.), *Artificial General Intelligence With 42 Figures and 16 Tables,* Springer-Verlag , Berlin – Heidelberg, 2007.

GUNKEL, D. J., *The Machine Question Critical Perspectives on AI, Robots, and Ethics*, The MIT Press, Cambridge, London 2012.

HOBBES, T., *Levijatan ili Građa, oblik i moć crkvene i građanske države*, Jesenski i Turk, Zagreb, 2004.

HOLLAND, J. H., *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, The MIT Press, USA, 1992.

HORST, S., *The Computational Theory of Mind*, The Metaphysics Research Lab Center for the Study of Language and Information Stanford University, 2003.

Pristupljeno 4.9.2019:https://www.researchgate.net/publication/277224413_The_Computational_Theory_of_Mind

JOHNSON, D.G., MILLER, K. W., Un-making artificial moral agents, *Ethics and Information Technology*, Vol. 10. Springer Netherlands, 2008., str. 123-133.

Pristupljeno 9.4.2019: https://link.springer.com/article/10.1007/s10676-008-9174-6

KOCH, C., *Consciousness Confessions of a Romantic Reductionist,* The MIT Press, Cambridge, 2012.

KURZWEIL, R., *The Singularity Is Near When Humans Transcend Biology*, Penguin Group, London, 2005.

MASAFUMI, O., ALBANTAKIS, L., GIULIO, T., From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0, *PLOS Computational Biology*, May 8 2014.

Pristupljeno 9.4.2019:https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003588

NIDA- RÜMELIN, J., *O ljudskoj slobodi,* Breza, Zagreb, 2007.

NIKOLIĆ, O., Utjelovljena svijest i naturalizirana fenomenologija, *Filozofska istraživanja*, (3/2017)

POOLE, D. L., MACKWORTH, A.K., *Artificial Intelligence Foundations of Computational Agents,* Cambridge University Press, New York, 2010.

Pristupljeno 4.9.2019: https://www.researchgate.net/publication/250333956_Robots_Should_Be_Slaves

REGAN, T,, A case for animal rights, *Advances in animal welfare science*, Washington, DC: The Humane Society of the United States. 1986/87., str.179-189.

Pristupljeno 4.9.2019: http://www.animal-rights-library.com/texts-m/regan03.pdf

REGAN, T., *The Case for Animal Rights*, University of California Press, 2004.

RUSSEL STUART, J., NORVIG, P., *Artificial Intelligence A Modern Approach*, Prentice Hall, Englewood Cliffs, New Jersey, Alan Apt, 1995.

SANDBERG, A., BOSTROM, N., *Whole Brain Emulation: A Roadmap*, Technical Report 2008-3, Future of Humanity Institute, Oxford University, 2008.

SEARLE, J. R., Minds, brains, and programs, *The Behavioral and Brain Sciences 3*, Cambridge University Press, Berkeley, California, 1980. str. 417-457.

Pristupljeno 4.9.2019: http://www.uh.edu/~garson/MindsBrainsandPrograms.pdf

SEIFERT, J.,*Das Leib-Seele Problem und die gegenwärtige philosophische Diskussion*, Wissenschaftliche Buchgesellschaft, Darmstadt, 1989.

SPARROW, R., Turing Triage Test*, Ethics and Information Technology* 6, Kluwer Academic Publishers Hingham, MA, USA 4/2004., str. 203–213.

Pristupljeno 9.4.2019: https://link.springer.com/article/10.1007%2Fs10676-004-6491-2

STEELS, L., BROOKS, R. (ed.), *The Artificial Life Route to Artificial Intelligence: Building Embodied,* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1995.

TONONI, G., An Information Integration Theory of Consciousness, *BMC Neuroscience* 5:42, 2004.

Pristupljeno 3.4.2019:https://bmcneurosci.biomedcentral.com/articles/10.1186/1471-2202-5-42

TORRANCE, S., Ethics and consciousness in artificial agents, *AI & Society* 22(4), April 2008., str. 495-521.

Pristupljeno13.4.2019:https://www.researchgate.net/publication/220415053_Ethics_and_consciousness_in_artificial_agents

TORRANCE, S., Artificial Consciousness and Artificial Ethics: Between Realism and Social Relationism, *Philosophy & Technology*, 27: 9, March 2014.

Pristupljeno 13.4.2019:
https://www.researchgate.net/publication/258168927_Artificial_Consciousness_and_Artificial_Ethics_Between_Realism_and_Social_Relationism

URBAN, T., *The AI Revolution: The Road to Superintelligence*, January 22, 2015.

Pristupljeno4.9.2019:https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html

VOLKERT, G., *Samoodređenje: princip individualnosti*, Demetra, Zagreb, 2003.

WARREN, M. A., *Moral Status Obligations to Persons and Other Living Things*, Clarendon Press, Oxford, 1997.

WEIDENFELD, N., Nida-Ruemelin, J., *Digitaler Humanismus: Eine Ethik für das Zeitalter der künstlichen Intelligenz*, Piper Verlag, München, 2018.

WEIDENFELD, O'CONNOR, T., WONG, H. Y., Emergent Properties, Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition).

Pristupljeno 7.4. 2019: https://plato.stanford.edu/entries/properties-emergent/

WEIZBAUM, J., *Die Macht der Computor und die Ohnmacht der Vernunft*, Frankfurt am Min, 1978.

WERNER, G., *Am Amfang war die Information*, Neuhausen, Stuttgart, 1994.,

WHITBY, B., Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents*, Interacting with Computers*, 20 (3). 2008., str. 326-333.

Pristupljeno16.4.2019:http://www.cs.potsdam.edu/faculty/laddbc/Teaching/Ethics/StudentPapers/2008Whitby-SometimesItsHardToBeARobot.pdf

WINSBY, M., *Suffering Subroutines: On the Humanity of Making a Computer that Feels Pain*, Western University, London, Association ofComputation and Philosophy Annual Meeting: College Park, MD 2013.

Pristupljeno4.9.2019:http://www.iacap.org/proceedings_IACAP13/paper_48.pdf

YUDKOWSKY, E., BOSTROM, N., „The Ethics of Artificial Intelligence",u: William Ramsey and Keith Frankish (ed.), Draft for Cambridge Handbook of Artificial Intelligence, Cambridge University Press, 2011.

Pristupljeno9.4.2019:https://nickbostrom.com/ethics/artificial-intelligence.pdf

ZAHAVI, D.,Intencionalnost i iskustvo, *Filozofska istraživanja* (2/2006)

Pristupljeno 16.4.2019:https://hrcak.srce.hr/search/?show=results&stype=1&c%5B0%5D=article_search&t%5B0%5D=Dan+Zahavi%2C+Intencionalnost+i+iskustvo

ZUCKERBERG, A. L., *Moral Agency and Advancements in Artificial Intelligence*, Claremont McKenna College, 2010.

Pristupljeno 4.9.2019:https://scholarship.claremont.edu/cmc_theses/36/

SAŽETAK: Etičko razmatranje moralnog statusa umjetne inteligencije

Ovaj rad se bavi raspravom oko moralnog statusa umjetne inteligencije, specifičnije, može li se umjetnu opću i super-inteligenciju smatrati moralnim pacijentom. Uvod u raspravu kreće od definiranja umjetne inteligencije, njezinih razina i strategija izgradnje umjetne opće inteligencije koje upućuju na kompleksnost navedenih sustava. Kako bi se razmotrila suvislost ove rasprave prvo se osvrće na uvjete pod kojima se umjetnu inteligenciju može smatrati moralnim pacijentom pa se potom dolazi do pitanja svijesti u umjetno inteligentnom sustavu te se ukratko razlaže komputacijska teorija uma, emergencija, kvantna teorija uma kao i integrirana teorija informacija Tononi Giulia i zaključci istraživanja neuralnih osnova svijesti Christofa Kocha kao svojevrsne replike na problematiku kineske sobe. Uz to se razlaže problematika dokazivanja svijesti kao iiskustva boli u umjetno inteligentnim sustavima. Status umjetne inteligencije se dovodi u konotaciju s pravima životinja te se primjenjuju određene verzije deontologije Toma Regana i utilitarizma Peter Singera te se razmatra u kolikoj mjeri su ove etičke teorije adekvatne pri raspravljanju o moralnom statusu umjetne inteligencije.

Ključne riječi: umjetna inteligencija, moralni status, problem svijesti, teorije svijesti, kineska soba, iskustvo boli u umjetnim sustavima, prava životinja, deontologija, utlitarizam

ABSTRACT: Ethical consideration of the moral status of artificial intelligent systems

This thesis is concerned with the discussion surrounding the moral status of artificial intelligence, specifically whether artificial general and superintelligence can be considered as a moral patient. Introduction into this discussion begins with defining artificial intelligence, its levels and strategies for construction of artificial general intelligence which points to the complexity of mentioned systems. In order to deliberate on the coherence of the whole discussion the text starts with the conditions needed to consider AI as a moral patient then it proceeds to problem of consciousness of artificial systems which is presented through short explanations of computational theory of mind, emergence, quantum theory of mind as well as Tononi Giulio's Integrated Information Theory and some conclusions Christof Koch's research regarding neural basis of consciousness as a retort to the problem of Chinese room. In addition, the problem of proving consciousness as well as the experience of pain of artificial systems is explained. The status of AI is brought toc onnection with animal rights and certain versions of Tom Regan's deontology and Peter Singer's utilitarianism are reviewed

and it is discussed to what extent they are adequate in application in the ethical discussion about the moral status of artificial intelligence.

Key words: artificial intelligence, moral status, hard problem of consciousness, theories of consciousness, Chinese room, experience of pain in artificial systems, animal rights, deontology, utilitarianism